Spatial visualization of uncertainty

SVEN CHRIST (BA Hons GEOM)

Thesis presented in partial fulfilment of the requirements for the degree of Master of Arts in Geography and Environmental Studies at Stellenbosch University

SUPERVISOR: MRS Z MUNCH

March 2017

DEPARTMENT OF GEOGRAPHY AND ENVIRONMENTAL STUDIES

DECLARATION

By submitting this research thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2017

ABSTRACT

Geospatial information has become more accessible since the early 2000s. Uncertainty has remained a constant in data, due to various factors, including scale and real world conceptualization. Geospatial products are frequently used to inform decision makers on key decisions, with little understanding of the quality of the data. However, accuracy assessments have improved significantly since the visual screening that was used in the 1950s, now providing statistics such as the Kappa coefficient, root mean square error (RMSE) and the confusion matrix.

Two questions thus arise: 1) do those using the data inform themselves about the quality of data; and 2) can visualization of the uncertainty in spatial data aid in the communication of the data quality? This research was achieved in three tasks: 1) evaluate the South African perception on data quality; 2) develop an uncertainty visualization tool; 3) evaluate the uncertainty visualization tool.

The first task was achieved through a quantitative survey of people working in the South African geospatial industry. Despite a limited response, the findings indicated that those working with geospatial data do not always seek to verify the quality of the data they are using. It also came to light that most of those who do not verify the quality of their data, would like to have the uncertainty in the data visualized.

Task 2 aimed at developing a tool for the visualization of spatial uncertainty (Uview). Uview was based on the findings from Task 1 supplemented by recommendations from literature and other uncertainty visualization tools. The tool was developed for continuous raster datasets only and uses the z-score and modified z-score as its main statistics for visualization. Standard accuracy assessment statistics (global data quality statistics), such as RMSE and mean absolute error (MAE) have also been included in Uview to make it an accuracy assessment and uncertainty visualization tool for continuous raster data.

Lastly Task 3, the evaluation of Uview was done using a two-pronged approach. The first part encompassed investigating the usability of the tool. In this phase the visualizations were used to derive relationships between digital elevation models (DEM), uncertainty and a watershed product. It was found that Uview does provide useful information, and watersheds are sensitive to deviations from true value at *key locations* more than the *magnitude* of the deviation.

When Uview was evaluated by twelve people in the geospatial industry they all agreed that though improvements can be made, as it presents itself currently it is already a useable product that can add value. All respondents agreed that the visualization improves the comprehension of the statistics, and so of uncertainty.

KEYWORDS

Uncertainty, uncertainty visualization, data quality, accuracy assessment, z-score, GIS, raster data

OPSOMMING

Ruimtelike inligting het sedert die vroeë 2000's meer toeganklik geword. Onsekerheid het 'n konstante in data gebly as gevolg van verskeie faktore, insluitend skaal en werklike wêreld konseptualisering. Ruimtelike produkte word dikwels gebruik om besluitnemers in te lig oor belangrike besluite, met min begrip van die kwaliteit van die data. Tog het akkuraatheid assessering aansienlik verbeter sedert die visuele metodes wat in die 1950's gebruik is, ook met die verskaffing van statistiek soos die Kappa-koëffisiënt, wortel-gemiddelde-kwadraat fout (RMSE) en die verwarringsmatriks.

Twee vrae ontstaan dus: 1) neem die gebruikers van die data die tyd om hulself te vergewis met die kwaliteit van data; en 2) kan visualisering van die onsekerheid in ruimtelike data die kommunikasie van die data kwaliteit ondersteun? Hierdie navorsing is behaal in drie take: 1) evalueer die Suid-Afrikaanse persepsie oor data kwaliteit; 2) ontwikkel 'n onsekerheid visualisering hulpmiddel; 3) evalueer die onsekerheid visualisering hulpmiddel.

Die eerste taak is behaal deur 'n kwantitatiewe opname van mense wat betrokke is in die ruimtelike inligtingsbedryf in Suid-Afrika. Ten spyte van 'n beperkte reaksie, het die bevindinge aangedui dat diegene wat met ruimtelike data omgaan nie altyd daarna streef om die data gehalte te verifieer nie. Dit het ook aan die lig gekom dat die meeste van diegene wat nie hul data gehalte verifieer nie, wel belangstel in ‘n onsekerheid visualisering van die data.

Taak 2 was gemik op die ontwikkeling van 'n instrument vir die visualisering van ruimtelike onsekerheid (Uview). Uview is gebaseer op die bevindinge van Taak 1 aangevul deur aanbevelings vanuit die literatuur en ander onsekerheid visualisering hulpmiddels. Die instrument is ontwikkel vir deurlopende roosterdatastelle en maak gebruik van die z-telling en gemodifiseerde z-telling as belangrikste statistieke vir visualisering. Standaard akkuraatheid assessering statistieke (globale data kwaliteit statistieke), soos RMSE en gemiddelde absolute fout (MAE) is ook ingesluit in Uview om dit 'n akkuraatheid assessering en onsekerheidsvisualisering hulpmiddel vir deurlopende roosterdata te maak.

Laastens die evaluering van Uview (Task 3) is gedoen met behulp van 'n tweeledige benadering. Die eerste deel het ondersoek ingestel na die bruikbaarheid van die instrument. In hierdie fase is die visualiserings gebruik om verhoudings tussen digitale elevasie modelle (DEM), onsekerheid en 'n waterskeiding produk af te lei. Daar is bevind dat Uview nuttige

inligting verskaf, en waterskeidings is meer sensitief vir afwykings van werklike waardes op belangrike plekke meer as die grootte van die afwyking.

Tydens die Uview evaluering deur twaalf mense vanuit die ruimtelike inligtingsbedryf, het almal saamgestem dat hoewel verbeteringe gemaak kan word, die produk soos dit tans daar uitsien alreeds 'n bruikbare produk is wat waarde kan toevoeg. Al die respondente het saamgestem dat die visualisering die begrip van die statistieke verbeter, en so ook van onsekerheid.

TREFWOORDE

Onsekerheid, onsekerheid visualisering, data kwaliteit, akkuraatheid assessering, z-telling, GIS, roosterdata

ACKNOWLEDGEMENTS

I would sincerely like to thank:

- My supervisor Mrs Z Munch, for her hours of support, guidance and assistance throughout the duration of this thesis;
- The NRF for their generous funding;
- Ms B Schöbel for more support than anyone can give, reading all my documents and making sure I get where I need to be;
- Mr T Sutton for assistance in learning the QGIS API;
- Mr G Ahnie for support, laughs and encouragement throughout;
- My mom for enabling me to dedicate extended time to this thesis;
- Mrs V Drotsky for always being there whenever I call;
- GISSA for helping me secure interviews;
- OSGEO for helping me secure interviews;
- All the respondents of my surveys and interviews;
- Mr B De Robeck for his editorial work;
- Everyone who took the time to listen even when my work sounded abstract.

CONTENTS

TABLES

FIGURES

ACRONYMS AND ABBREVIATIONS

API	application programming interface
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
BoK	Geographic Information System and Technology Body of Knowledge
CGA	Centre for Geographic Analysis
DEM	digital elevation model
EC	European Commission
ERC	European Research Council
GIS	geographic information system
GISc	geopgraphic information science
GISs	geographic information systems
GISSA	Geo-Information Society of South Africa
LAS	LASer
LiDAR	light detection and ranging
LP	layers panel
MAD	median absolute deviation
MAE	mean absolute error
OSGeo	Open Source Geospatial Foundation's
OVI	overall visualization index
PLATO	Council for Professional and Technical Surveyors
REC	Stellenbosch Research Ethics Committee
RMSE	root mean square error
ROC	receiver operating characteristic
SAQA	South African Qualifications Authority
SD	standard deviation
SRTM	Shuttle Radar Topography Mission

SUDEM	Stellenbosch University Digital Elevation Model

CHAPTER 1 A VIEW INTO GEOSPATIAL UNCERTAINTY

One of the key functions of a geographic information system (GIS) is that it enables new information to be derived from spatial data files already held, such as gradient and aspect from digital elevation models (De Gennaro et al. 2014; Longley et al. 1999). This is especially useful in environmental and earth sciences where the resultant datasets and models are often treated as completely accurate and used with absolute confidence (Longley et al. 1999; Goodchild 1996). However, the existence of error is always a factor when dealing with spatial data (Wong & Sun 2013; Jacquez 2012; Couclelis 2003; MacEachren 1992). This error factor can be referred to as 'uncertainty', making it good practice to always, after any modelling, perform an accuracy assessment to measure the difference between actual reality and the representation's notion of reality. 'Uncertainty' however is not a simple concept, MacEachren et al. (2005:140) describes uncertainty as: "when inaccuracy is known it can be defined as error; when it is not known, the term uncertainty applies." Longley et al. (2005:100) describes uncertainty as "the difference between the contents of the dataset and the phenomena that the data are supposed to represent."

Before data can be useful in a GIS it needs to undergo transformations. These transformations create at least three opportunities for uncertainty to enter the data: 1) during real world to a human conception of the world; 2) when this human conception is measured with some device; and 3) during analysis of this measurement (Longley et al. 2005). Each of these steps represents a transformation of the real world which could affect the eventual spatial representation. Therefore, all datasets come with inherent uncertainty from a possible multiple range of inputs. Longley et al. (2005) therefore suggests using a fuzzy [logic] approach when capturing data from the real world to analysis, thereby representing 'degrees of truth'. Foody (2002) stated that errors are a part of maps and spatial data, as they are merely a generalization of the world.

The fact that no piece of data is completely error-free, be it actual error or because of statistical variation, suggests that uncertainty will always form part of analysis (Longley et al. 1999). In addition, when multiple datasets are combined to create a product, uncertainty and error can propagate from imperfect input data to the final output. This is especially relevant when the data output of one process is the data input into another (Longley et al. 1999).

Accuracy (Merriam-Webster s.a.) is defined as the "degree of conformity of a measure to a standard or a true value." In recent years, there have been great advancements in accuracy

assessment techniques (Pontius & Millones 2011; Lunetta & Lyon 2004). Accuracy assessment is no longer an afterthought, but has become a key feature of GIS datasets (Foody 2002). Whereas in the past a visual inspection may have been adequate, now a more scientific and statistically based technique is often required. Methods such as the Kappa coefficient and, in particular, the confusion matrix are used to define overall accuracy of a product. These tell the user the percentage of agreement between product and reference data (Foody 2002). Root-mean-square-error (RMSE) is another commonly used assessment of accuracy (Aguilar, Agüera & Aguilar 2007). One method that does produce visual output is the receiver operating characteristic (ROC) analysis, which can be used to assess binary classification models for both rank order and continuous datasets (Mas et al. 2013). ROC is a graph plot of the probability of having a true positive versus a false positive, as the probability cut-off varies (Feizizadeh, Jankowski & Blaschke 2014). Most of these methods are however statistical, representing numeric values, but they do not represent visually where the uncertainty may occur.

Because visualization trumps text in its impact and appeal, one has to look at cartography, a field closely related to GIS, that largely focusses visual communication of spatial results for potential techniques (Bostrom, Anselin & Farris 2008; MacEachren 1992). Whilst many studies on representing uncertainty have been undertaken in the field of cartography (Kinkeldey, MacEachren & Schiewe 2014; MacEachren et al. 2005; Howard & MacEachren 1996), with its strong focus and history on data quality, a classic example of uncertainty visualization is the bivariate representation proposed by Howard & MacEachren (1996) in R-Vis which will be discussed in detail in Chapters 2-4.

Since the models in GIS and spatial modelling are closely linked with visualization, one can also use models to represent uncertainty, such as epsilon bands (Petrasova et al. 2014; Shi, Fisher & Goodchild 2002; Bishop & Karadaglis 1996; Fisher 1995). Though other methods exist, recent research has indicated that the use of uncertainty visualization is closely linked to user and task abilities and requirements (Kinkeldey, MacEachren & Schiewe 2014).

A key area where uncertainty visualization may also play a critical role in environmental studies, is in spatiotemporal change analysis. When old paper maps and photographs are digitized, the inherent uncertainty in these old datasets, as well as in the conversion process, may play a role in the eventual findings of the study (Jenny & Hurni 2011).

1.1 REAL WORLD PROBLEM

Since the inception of GIS there has been an increasing uptake in the use of spatial data. In the decades of the early 1960s and 1970s GIS was limited to experts in the field of geography, mathematics and computer science (Hessler 2014). Today, with the widespread availability of open source products such as QGIS previously known as Quantum GIS (QGIS s.a.a), GIS is available to anyone who wishes to spend the time performing a wide range of applications using these tools. Then there are the well-known Google Maps (Google Maps s.a.) and Google Earth (Google Earth s.a.) products, which both hold geospatial data and GIS functionality. Spatial datasets for most countries can be downloaded from various online sources which capture data through varying methods, but do not supply supporting metadata or projection information. Often only a text file exonerating the supplier from any liability arising out of using the data is found, excluding also any indication of the scale or level of accuracy of the data (MapCruzin s.a.). Even if supplied, metadata does not always provide the data lineage record of the life cycle of that data that indicates how it was created and what in it has been altered. It must be noted that a lack of metadata is not only a shortcoming of data suppliers, but also of GIS users who do not always actively look for it. This leads both to users not fully understanding the quality of the data they are using and to an increased risk of analysis error.

Less experienced GIS users may treat data as perfectly accurate and incorporate it into a multitude of uses ranging from scientific analyses to governmental decision making and planning (Van Oort 2005). The end user who receives the finished product is often even less aware of the uncertainty in both the data product and in the processing (analysis) chain (Van Oort 2005). Users who are aware of the uncertainty factor develop different strategies for dealing with uncertainty; more experienced users may consider statistics but ignore the spatial component, thereby discarding potentially useful data. On the other hand, there are those who simply rely upon heuristic discovery methods which, whilst practical, are not always reliable or of optimal value. Still worse, some users just ignore the uncertainty of data completely, with increased potentially harmful results (Monmonier 2006; MacEachren et al. 2005).

One could thus deduce that the availability of GIS software and datasets creates a risk of producing output that cannot and should not be accepted as perfectly accurate without proper understanding and documentation. The use of datasets without proper metadata, as well as end users without proper knowledge of the accuracy of their data, leaves a gap where

uncertainty is not communicated. Even if metadata is supplied, it may not be spatially clear where uncertainty occurs in that data, with users therefore at risk of not understanding the quality of the products derived and the quality of the decisions that can be taken with such data. This lack of understanding and uniform processing of uncertainty leads to the formulation of the research problem.

1.2 RESEARCH PROBLEM

In an ideal representation of the world every piece of data would be 100% accurate and free from error. In reality however, this is not the case; statistics have been used to represent uncertainty, whilst visualization, although researched, has been neglected. The effects of uncertainty can be widespread, as spatial representation in GIS is often used as a decision-making tool (MacEachren et al. 2005). Visual representation of uncertainty in spatial datasets could therefore provide users with greater confidence, not only in the spatial data products that they use, but also in the decision-making processes that are reliant on such data.

The following research questions therefore arise:

- Can uncertainty which is inherent in spatial data or produced through modelling be visualized to facilitate understanding of the data quality by those who use and produce spatial data?
- How does visualization of uncertainty affect users' perception of spatial products produced?

In answering these two questions, research into uncertainty visualization is a critically necessity in order to enhance the visualization of uncertainty and make the understanding of data quality easier to non-expert and amateur users. As continuous data is used for environmental data, that affects both environmental and societal development combined with the differences between continuous and discreet data structures, continuous data was selected.

1.3 RESEARCH AIM AND OBJECTIVES

The problem of uncertainty in spatial representation is not always made clear to decision makers and data users who are of varying levels of expertise in interpreting and using spatial data. The following aims and objectives have been set to resolve the problem:

1.3.1 Aims

The aims of this research study are therefore to:

a) determine the extent that users of spatial data understand the quality of their data; and

b) develop a tool that could aid in communicating the overall quality of datasets.

1.3.2 Objectives

1. Establish a baseline perception of general data quality amongst users and producers when working with spatial data;
2. Evaluate available uncertainty visualization tools for raster data through literature;
3. Develop software tool for uncertainty visualization of continuous raster data;
4. Generate visualization scenarios to test the software tool;
5. Compare the effect of statistical and visualized uncertainty on the perception of users and producers of spatial data, as well as on their decision making.

In order to demonstrate achievement of these objectives, a modelled dataset within the following study areas has been chosen, to best facilitate the study.

1.4 STUDY AREAS

Two study areas have been selected in the City of Cape Town. One on the sea shore and the other in the Helderberg basin. Two digital elevation models (DEMs) were chosen: 1) the Stellenbosch University Digital Elevation Model (SUDEM) at 5 m resolution; and 2) a digital elevation model (DEM) created from LASer (LAS) files resampled to 5 m (explained below).

A DEM is a continuous representation of elevation values over a topographic surface using a regular array of z-values (ESRI s.a.a). The SUDEM, an output of spatial modelling, was created through a fusion method that included datasets such as 20 m contours and, where available, 5 m intervals vertical interval contours; spot heights were also captured from 1:10 000 orthophotos. The 1:50 000 orthophoto series was only used in areas where large scale data was not available. The Shuttle Radar Topography Mission (SRTM) 'research-grade' version was also used in combination with the contours and elevation points in somewhat flat areas. For accuracy assessment, the mean absolute error (MAE) and RMSE of the fused DEM was calculated using airborne light detection and ranging (LiDAR) points at centimetre resolution (Van Niekerk 2012).

LAS is a format of sequential binary storage, used in this case to store data from a LiDAR laser. LAS files are natively stored as dense point clouds. The LAS files were processed by

the Centre for Geographic Analysis (CGA), using the procedure documented by Fagan & Maidment (s.a.), to produce a 5 m DEM. Overall quality of LiDAR datasets is largely dependent on quality of the calibration of the LiDAR system (United States Geological Survey 2014). LiDAR systems consist of a laser ranging and scanning unit in tandem with a position orientation system that encompasses a global navigation satellite system and an inertial navigation system. A LiDAR system returns a high density 3D point cloud, whose neighbouring points from multiple swaths can be used to gather inter-swath quality (United States Geological Survey 2014; Habib & Van Rens 2008). As LiDAR data is a high point cloud system, it is more likely to both get true ground readings and not be affected by shadow such as stereo imaging techniques (FUGRO s.a.). This explains why the LiDAR dataset was chosen as a validation dataset to the SUDEM.

The areas selected were covered by both the SUDEM and the LiDAR dataset. To evaluate the accuracy of the SUDEM and determine the degree of uncertainty, a high resolution LiDAR dataset was chosen.

The two selected study areas are highlighted in Figure 1.1 below. The sea shore study area is the one used for development of the software tool, whilst the Helderberg basin study area is used for its evaluation.

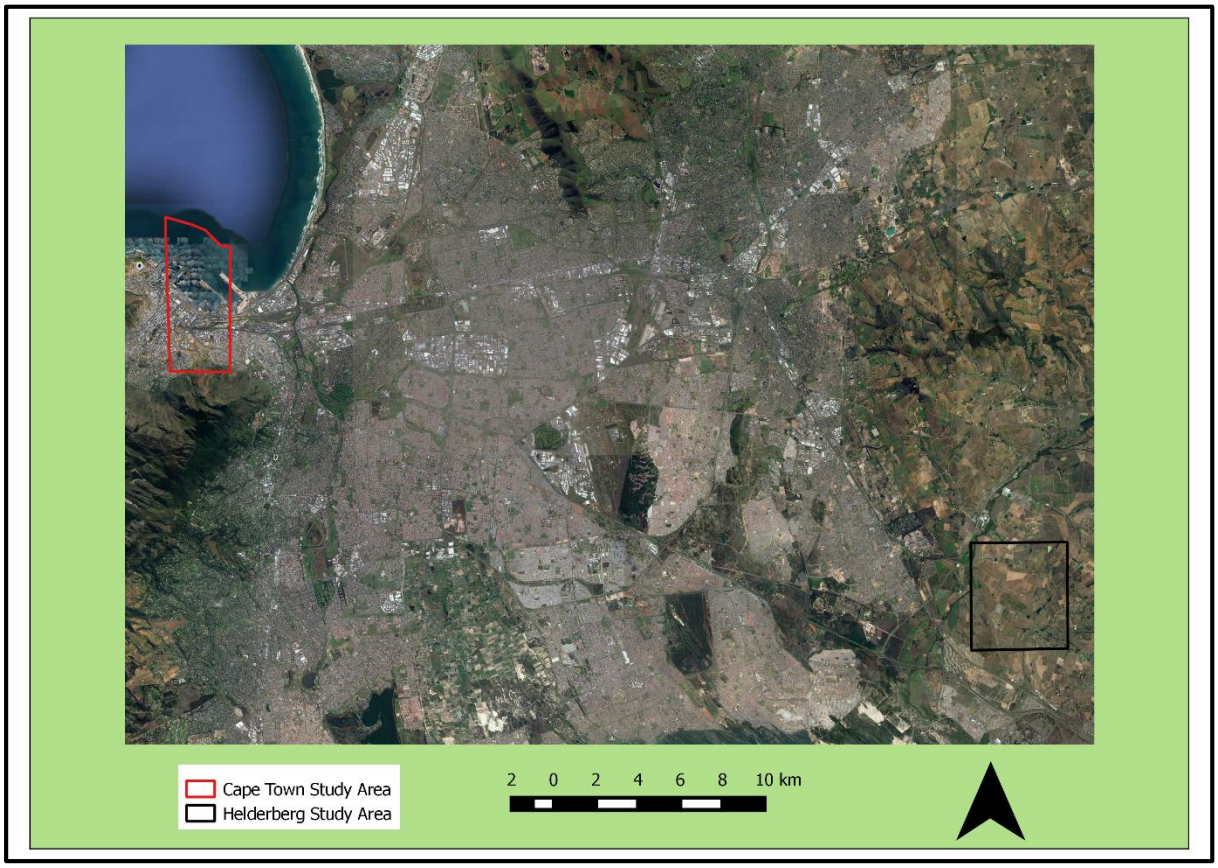


Figure 1.1 Study area

These two areas were selected as they both contain sharp elevation fluctuation coming up from low land to a mountain, with some places being near cliff faces. This topography causes geospatial uncertainty as these sharp elevation changes have to be merged into 5 m blocks. Points will be sampled (collecting elevation information) from the LiDAR dataset and compared to the SUDEM in order to create an uncertainty map using the developed tool.

1.5 METHODOLOGY AND DESIGN

This study uses both qualitative and quantitative methods. The hypothesis put forward is that users and producers of spatial data are conscious of the quality of their data. This research has been broken down into three main tasks, with linked subtasks (see Figure 1.2). The first two tasks are based on quantitative statistical methods, whilst the third is based on quantitative

methods for the simulation and watershed evaluation and qualitative methods for the interviewing of key personnel involved with using and producing spatial data.

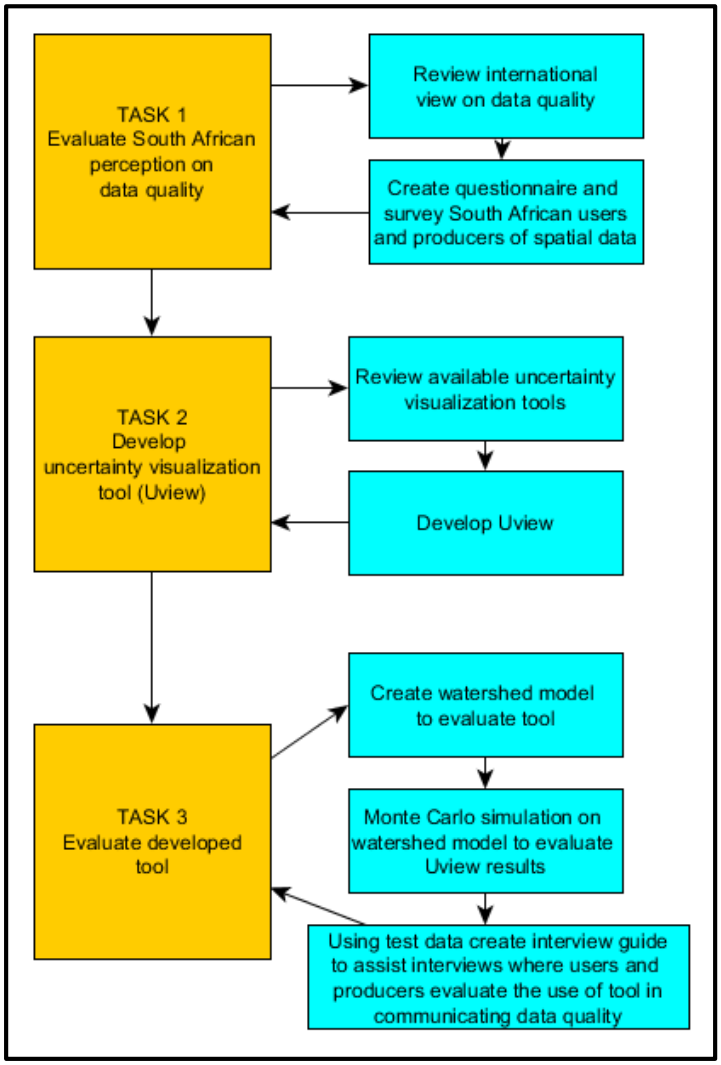


Figure 1.2 Research tasks

Task 1 addresses the first objective, namely to establish the baseline perceptions amongst users and producers. It contains a synopsis of the international view on data quality, achieved through viewing works such as those by Kinkeldey & Schiewe (2014), De Graaff (2013) and Alberti (2013). This synopsis then informed the choice of questions to be put to South African users and producers, so as to create the baseline of what the sentiment towards data quality is in South Africa. The survey also informed the selection of visualization techniques chosen in the next phase of the research.

Task 2 addressed the second and third objectives of software tool evaluation and development for raster data uncertainty visualization. It followed a similar protocol to Task 1, where the first subtask was to establish what is already available in literature and practice.

This information, together with the information from Task 1, was then used to design and develop the software tool. Cartographic principles of communication were used to increase the usability and understanding of the resultant uncertainty maps. Available information from tools such as *Aguila* (Pebesma, De Jong & Bierkens 2007) and *UncertWeb* (UncertWeb s.a.) was also used to inform the development of the software tool. Research by Kinkeldey & Schiewe (2014) indicated if an uncertainty visualization tool is to be created, it should be either for QGIS or ArcMap. Thus QGIS, as one of the leading open source GIS software products, was chosen as the basis to enable the tool developed in this study freely available to all, by being a QGIS plugin. The researcher also had some prior knowledge of both QGIS and Python, the language used for QGIS plugins.

Evaluation of the developed tool was the purpose of Task 3, which addressed the creation of usable results from Uview and then evaluated the use and power of Uview. Two watershed models were then created from the SUDEM and LiDAR dataset at 5 m resolution, with the results comparison based on the uncertainty map created from the study software tool. This served as an internal validation of the study tool, and as input for the third subtask. The second subtask Monte Carlo simulation of the watershed was again used to evaluate how the areas of high uncertainty, based on Uview, correlated with the modelled data and relate to Uview's findings.

The third subtask entailed qualitative interviewing of users and producers of spatial data; a demonstration of the developed study tool was given using data from subtask one, followed by a discussion on how this affected the sentiment towards the data and whether the tool was effective. Another test was then done, based on the two watershed models, whereby the SUDEM was corrected at the points of most uncertainty between the datasets as indicated by Uview, and the watershed model was run again on this partially corrected SUDEM. The results were then compared to the watershed output of the LiDAR dataset and visually compared with the watersheds for the original SUDEM and LiDAR datasets.

The methodology for each task will be fully defined and explained in the applicable chapters.

1.6 DATA SOURCES

Continuous raster datasets were chosen as the focus of the study tool, due to these types of datasets often being used for environmental data. The main literature sources were Google Scholar (Google Scholar) and the Stellenbosch University Library's (Stellenbosch University Library and Information Service) online database search tool and hard copy book index.

A working model of Alberti's 2013 work was also received through personal communication with him. Further, the Geo-Information Society of South Africa (GISSA) and the Open Source Geospatial Foundation's (OSGEO) Africa forum was used to contact parties interested in the study.

The two datasets were both received from the CGA, with the LiDAR data belonging to the City of Cape Town, but processed by the CGA, the SUDEM being a product of the CGA.

1.7 SIGNIFICANCE OF RESEARCH

GIS use in the researchers' enquiry has highlighted the difficulty that uncertainty holds in data. From a research perspective, work may be questioned in terms of the certainty of the dataset used, especially when fellow researchers ask questions at conferences. It would be best if one could have a good, solid visual answer for such questions.

1.8 LIMITATIONS

This study is limited to the use of continuous raster data, as raster data and vector data differ in terms of their structure. Raster files store data in grid blocks, whilst vector datasets store data as points, lines and polygons (Neumann, Freimark & Wehrle 2010).

Raster datasets store data in the same way as JPEG images from any 'point-and-shoot' camera. These datasets are made of grid blocks called pixels, each of which contains a single value. Vector data, being lines, points and polygons, stores information similar to basic stick drawings. These vector elements can, however, have multiple attributes ascribed to them (Kraak & Ormelling 2011).

Raster datasets were chosen due to the above mentioned differences in data structure and coding of data, but further as they are more frequently used in spatial modelling.

Due to time constraints and difficulty in obtaining copies of most of the currently existing tools for uncertainty visualization, as well as complications installing the tools that are available, tools were reviewed in literature only. Thus, lessons learned were mostly limited to theoretical points.

1.9 STRUCTURE OF THE PAPER

Chapter 1 provides an introduction to geospatial data uncertainty, visualization and measurement need and problems, the study research aims and objectives, data sources and overall study limitations. In Chapter 2 relevant literature is reviewed, which both informs the

research as to what knowledge is already available and further guides the research in structure and academic worth. The survey in Chapter 3 gives a view into uncertainty in South Africa, further informing readers of the needs of South Africans and if a tool for visualization would be useful for this audience. Chapter 4 starts off with an evaluation of existing uncertainty visualization tools. From this, a framework for Uview is development, based on the findings from Chapter 3 and the software evaluation. The development of Uview is then documented along with the metrics used for uncertainty visualization. Chapter 5 focusses on the use and evaluation of Uview through models and Monte Carlo simulation. The chapter starts with introducing the reader to the installation and use of Uview followed by an evaluation of Uview. The evaluation is done through the use of simulations and the comparing of watershed models between the SUDEM and the LiDAR dataset with the Uview product. In Chapter 6 Uview is evaluated through the use of a selected group of interviewees. This is a qualitative review to discover if Uview is useful as well as what the short comings are, and what further improvements are needed. Chapter 7 is a review of the knowledge gained through the research and conclusions drawn.

CHAPTER 2 UNCERTAINTY, DATA QUALITY AND CARTOGRAPHY

This chapter serves as the literature review, with an introduction of key topics of the research. An introduction to uncertainty is followed by investigating the academic knowledge requirements around uncertainty, what Geographic Information Science and Technology Body of Knowledge (BoK) states about uncertainty and how that links up with the South African Council for Professional and Technical Surveyors (PLATO) regulations for registration as a geographic information science (GISc) professional. Following this, the international view of uncertainty is examined to familiarise the reader and be able to work with what is understood by uncertainty. Goevisualization as a method of communication is investigated before the focus is turned to visualization specific to uncertainty. Different methods of visualization of uncertainty are then investigated that are currently available. Raster and vector data are examined for strengths and weaknesses, as both are common geospatial data formats and the visualizations that can be applied to them; the main focus in this study is on raster data therefore an in depth look into raster data visualization is given. The focus in this chapter is then turned to accuracy assessments and their relation to visualization. A brief introduction is given to colour representation, as colour is a fundamental aspect of visualization and it will be discussed how colour works and helps understanding of data. After the investigation of colour, a software development framework for visualization tools is investigated, leading to the final evaluation of what software would be best to develop an uncertainty visualization tool in.

GIS excels in its ability to derive new information from datasets already held, this is one of the most important aspects of GIS, however this is also why uncertainty management is such a crucial element as error is always a factor when dealing with spatial data (De Gennaro et al. 2014; Wong & Sun, 2013; Jacquez 2012; Longley et al. 1999; MacEachren 1992). The derived datasets such as slope and aspect from a digital elevation model (DEM), are often perceived as highly accurate without considering the documentation it may be accompanied by or the quality of the original DEM (De Gennaro et al. 2014; Longley et al. 1999). For this reason accuracy assessments have become a key element in dataset creation and data modelling (Lunetta & Lyon 2005).

Accuracy assessment has come a long way since its start in the late 1950s to early 1980s, when it was frequently considered an afterthought. In those early days, a visual inspection

was often used to determine if a product was acceptable or not (Ross & Lyon 2005; Foody 2002). Today however, more effective methods of assessment are being used. Kappa coefficient and, in particular, the confusion matrix are used to define overall classification accuracy of a product. These tell the user the percentage of agreement between product and reference data, although some criticisms have been raised around the efficacy of Kappa, especially relating to chance agreements (Foody 2002). Root mean square error (RMSE) is another common assessment of accuracy (McNyset, Volk & Jordan 2015; Chai & Draxler 2014; Aguilar, Agüera & Aguilar 2007). However, these are statistical methods represented by using numeric values; they do not represent visually where the error may occur. A lot of this work regarding visualization of uncertainty has been done in a field closely related to GIS named cartography, which has a strong focus and history on data quality (MacEachren 1992).

Currently more time is spent in the determination of error and uncertainty (accuracy assessments) than in taking the original measurements, or in this case, making the original datasets (Nondestructive Testing Resource Centre s.a.; Smits, Dellepiane & Schowengerdt 2010). Therefore the value of this costly time intensive exercise should not be hidden or lost without being communicated. For visual communication of uncertainty, MacEachren is identified as a key figure, not only in uncertainty visualization, but in general cartographic visualization. In 2012 the European Research Council (ERC) advanced a grant award to Prof. Dr. Rüdiger Westermann, with a value of 2.3 million Euros, for research into uncertainty visualization, which indicates the relevance and importance of research in the field of visualization (Technische Universität München s.a.). Other works such as Slocum et al. (2013) also indicate visualization as an area that has received attention lately, but which continue to require more research.

The next section will introduce the reader to the concept of uncertainty or rather the uncertainty around the concept of uncertainty, before going into uncertainty visualization, ending with the most useful software for which to create an uncertainty visualization tool.

2.1 UNCERTAINTY OR PROBABILITY

In GIS, two of the more common distinctions of uncertainty are those by MacEachren et al. (2005) and Longley et al. (2005). MacEachren (2005) holds uncertainty as inaccuracy that cannot be measured, whereas Longley (2005) describes uncertainty as the difference between the content of the dataset and the data that it is supposed to represent. MacEachren et al.

(2005:140) describes uncertainty as "when inaccuracy is known it can be defined as error; when it is not known, the term uncertainty applies." Longley et al. (2005:100) describes uncertainty as "the difference between the contents of the dataset and the phenomena that the data are supposed to represent." Therefore in Longley's description uncertainty can be quantified whereas for the MacEachren's definition quantifiability is problematic.

Uncertainty has also been defined as being either statistical variation or spread, error and maximum-minimum ranges, without making distinguishing between error and uncertainty (Wittenbrink et al. 1996). Mowrer (2000) and Foody & Atkinson (2003) both however, refer to uncertainty as something that can be quantified and stated about the correctness of a point.

In this thesis for the definition of uncertainty, a combination of these will be accepted. Uncertainty is the difference between a dataset and what it represents, be it statistical variation or error, further it can be expressed through the use of statistics and extended to areas close to the point measured.

Uncertainty and probability are often related aspects. Probability can be used to decide if useful information can be derived from incomplete or uncertain data (Candes, Romberg & Tao 2006). Probability however is almost as contentious as uncertainty. There are two schools of thought on probability, one by Subjectivists and the other by Frequentists. Subjectivists relate probability to the belief that is held by an entity, that an event may occur, whilst Frequentists believe, it is the frequency at which an event may occur that determines probability (Miller & Childers 2004). Probability and statistics can thus be combined to treat uncertainty as degrees of confirmation or strengths of belonging.

In relation to probability, one could endeavour to define uncertainty to be linked to the belief (trust) one has in the quality of the dataset, i.e. the lower the trust one has in the data, the higher the uncertainty. For the creation of the overall quality report (accuracy assessment) the Frequentists' belief that the frequency that the dataset has been correct or how close to correct the dataset is, relates to how good the data quality is. However the accuracy assessment is read as the belief (confidence) one has that the dataset overall is of an acceptable quality.

2.2 ACADEMIC REQUIREMENTS FOR KNOWLEDGE ON UNCERTAINTY

GIS under the South African Geomatics Professional Act, Act No. 19 of 2013, has become a protected profession. Like becoming a doctor or lawyer, anyone wishing to practice as a GIS professional in South Africa is required to be registered (Rethman 2014). In South Africa, the

body that registers GISc practitioners, is the South African Geomatics Council formerly known as the South African Council for Professional and Technical Surveyors (PLATO). In this thesis, the body will however still be referred to as PLATO, as all literature used still use the name PLATO, further the same academic model and registration process applies.

An accreditation and professional registration board is responsible for safeguarding and promoting the professional interests of the profession, as well as fostering high educational standards for education and professional practice (Jefferies & Evetts 2000).

Accreditation by professional bodies is a method of giving a certain legitimacy and assurances to the quality of qualifications given by educational institutions (Du Plessis 2015). This in itself can be misleading, as accreditation does not necessarily ensure quality or understanding of key concepts (Beaulieu & Epstein 2002). PLATO thus has direct input into what should be taught at universities and what should be known by students through their required academic accreditation model.

PLATO has drawn their requirements for accredited GIS courses from many sources. The Geographical Information Science and Technology Body of Knowledge (BoK) is a piece of literature that covers a broad spectrum of knowledge, drawn from international standards, which can and should be used as a reference point for the requirements for GIS/Sc programmes, it is frequently used internationally for curriculum development (Du Plessis 2015; Du Plessis & Van Niekerk 2014). It is however not completely comprehensive; further documents should be consulted in the process of curriculum development (Du Plessis 2015). The PhD by Heinrich Du Plessis (2014 vice president of PLATO) on accreditation and an academic model for South African GIS accreditation makes mention of the South African Qualification Authority (SAQA), the BoK and PLATO all referring to uncertainty and/or data quality as being a key field of knowledge that must be attainted. However, the form in which it appears in the dissertation, gives little clarity on what this entails. The PLATO guideline for professional GISc for 2015, mentions the knowledge of the impact of uncertainty once, under a category of map use and evaluation, which consists of eight (8) credits (PLATO 2015). This equates to 1.66% of the total credits required for registration as a GISc professional. Another eight (8) credit requirement is mentioned for data quality. This section deals with primary and secondary data as well as data accuracy, completeness and metadata. This is another 1.66% leading to a total of 3.32% of the required educational time being spent on uncertainty and data quality. With data quality being at the core of usable results, the question is whether this is enough in reality. The BoK and PLATO model differ further in

that mathematical sciences and physics are required for PLATO, but are not a key knowledge area in the BoK (Du Plessis & Van Niekerk 2012), whilst the BoK mentions data quality and uncertainty in no less than three (3) areas in a theoretical context as well as in an analysis context (DiBiase et al. 2006). The BoK includes uncertainty from the level of identification in different contexts, to the mathematical uncertainty models relating to statistics and probability (DiBiase et al. 2006). The BoK further includes uncertainty visualization under the knowledge area of analytical techniques (DiBiase et al. 2006).

It is therefore notable that the international BoK takes a wider stance on the knowledge of data quality than the South African PLATO model. The BoK also goes into great detail of what is included under knowledge of data quality (DiBiase et al. 2006), whereas the PLATO model is unclear as to what knowledge is required under data quality. In the South African context then, there is scope and awareness but not a broad enough teaching of the knowledge of uncertainty and management thereof. It further is important to note that producers of geospatial data are regulated, but not end users and decision makers.

2.3 INTERNATIONAL UNDERSTANDING ON UNCERTAINTY AND VISUALIZATION

Uncertainty is present, to a greater or lesser extent, in all datasets worldwide, making it essential to look into what is understood internationally. De Graaff (2013) found that there are different definitions and understandings of uncertainty based on user experience. Some professionals view high spatial data quality as referring to ‘data that is free of error’. It was also found that different users require different methods of uncertainty representation, be they statistical or visual. A discrepancy was found between how expert and novice users perceive uncertainty. With a further complication found in that not all providers provided sufficient information about data quality (metadata) (De Graaff 2013).

Other studies indicate that uncertainty is regarded as a fuzzy concept (Kinkeldey & Schiewe 2014). Participants in this study regarded uncertainty visualization to be especially relevant in change analysis by means of remote sensing (Kinkeldey & Schiewe 2014). Where areas of change fall in areas of higher uncertainty further, investigation can improve the classification, thus improving the quality of the change product. This thus agrees with Zhang & Goodchild (2002), that visualization can improve the identification of information about uncertainty, uncertainty trends and knowledge on where data should be improved. Respondents from the study of Kinkeldey & Schiewe (2014) also claimed that they can thus have more confidence

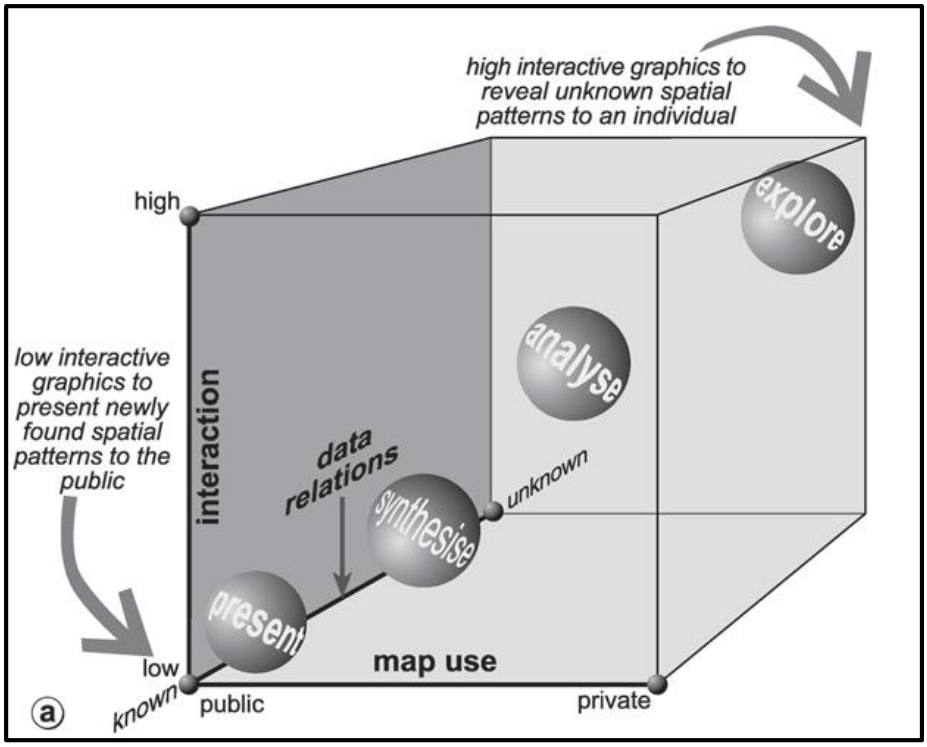
in their work, as well as better understanding of the quality of the data, if uncertainty is visualized. Out of the three groups of experts interviewed by Kinkeldey & Schiewe (2014), one group felt that uncertainty should be visualized for decision makers, the second group was unsure if such information would be valuable to decision makers, whilst the third group felt that this information may degrade the way decision makers feel about data. All three groups agreed that visualization is a powerful method of communication. These experts also felt that, if a tool were to be developed for uncertainty visualization, it should be a plugin for commonly used GIS software such as ArcMap and QGIS (Kinkeldey et al. 2015; Kinkeldey & Schiewe 2014). ArcMap and QGIS will be further discussed in section 2.10 [Common GIS software].

Research into uncertainty and data quality uses and understanding was done by Tegtmeier et al. (2007). Twelve questionnaires were submitted to Dutch civil engineering companies, in order to ascertain how they treat uncertainty information and how they may deal with it. Three of the twelve companies (25%) indicated, that they pay no attention to uncertainty information at all (Tegtmeier et al. 2007). Seven companies (58%) had an inspection of quality done at representation level, with usually only a supervisor or co-worker evaluating the work visually and without the aid of computers or statistics. Only one company (8%), which specialises in visualization, used computer software to understand and visualize their data quality (Tegtmeier et al. 2007). The last response was not mentioned, the findings however also referred to expert influence of data (conceptualizing of real world) being taken into account as a potential source of uncertainty by two companies (Tegtmeier et al. 2007).

What this research indicates is that different users have different needs, be they visual or statistical. Further it can be deduced from literature, that there is neither a uniform definition of uncertainty in GIS nor a uniform understanding (Kinkeldey & Schiewe 2014; De Graaff 2013; Tegtmeier et al. 2007; Longley et al. 2005; MacEachren et al. 2005). It also identified that visual representation of uncertainty could be seen by some as degrading the perceived value of a product and not ideal to show to decision makers, but can also aid in spatial data exploration. In cases where uncertainty is visualized, care should be taken as to how colour is used. Nonetheless, all studies indicate that visualization is a powerful tool that can aid communicating uncertainty in a powerful fashion.

2.4 GEOVISUALIZATION

As visualization has been identified as an area that can improve the understanding of uncertainty, geovisualization should be addressed as a concept (Kinkeldey et al. 2015; Tegtmeier et al. 2007; Zhang & Goodchild 2002). Geovisualization has its roots in scientific visualization, a set of techniques developed outside geography. Such techniques were mostly visualization of medical imaging, molecular structure and fluid flows (Slocum et al. 2013). The aim was to use existing scientific information to create new insight through visual methods. That geovisualization has a similar purpose can be seen by the definitions of MacEachren (1998). Firstly, geovisualization is the use of interactive or static maps to visually show spatial contexts and problems visibly, in order to use the most powerful human information processing abilities, namely those related to vision (Slocum et al. 2013; Howard & MacEachren 1996). Secondly, MacEachren (1998) used a cube shaped diagram to put visualization and communication on opposing sides of the cube, indicating that they are the two extremes of any visualization (see Figure 2.1). Geovisualization can also be understood as a visual method to facilitate geographic thought (Howard & MacEachren 1996). MacEachren however, maintained that some maps represent features of both visualization and communication. Before attention can be turned to the Visualization vs. Communication Cube, one must first understand what MacEachren understands as communication. He describes communication as a public activity in which knowns are presented in a non-interactive environment. By contrast, he describes visualization as a private activity that is highly interactive where unknowns are revealed (Slocum et al. 2013; MacEachren 1998). Understanding this, the focus can now be turned to the cube in Figure 2.1.



Source: adapted from MacEachren & Taylor 1994

Figure 2.1 Visualization vs. Communication cube

What the cube communicates is that visualization is not just the representation of data that is easily interpreted, but rather that it is an interactive process through which new findings are discovered (MacEachren 1998; Slocum et al. 2013). The cube however indicates that a map may be a combination of both (visualization and communication) or solely a visualization or communicative device (MacEachren 1998). A 'you are here map' such as those indicating one's location in a shopping mall, provides immediately accessible data and is therefore a communicative device, whereas a map representing the change over time of land cover, would not only require thought for it to be comprehended, but also could bring new information to light as to land cover change patterns.

To sum up, geovisualization is the process of viewing a map, interpreting the image and gaining new information and insight from it. It is not just a representation of obvious data, but an interactive process that brings new information to light.

2.5 UNCERTAINTY VISUALIZATION

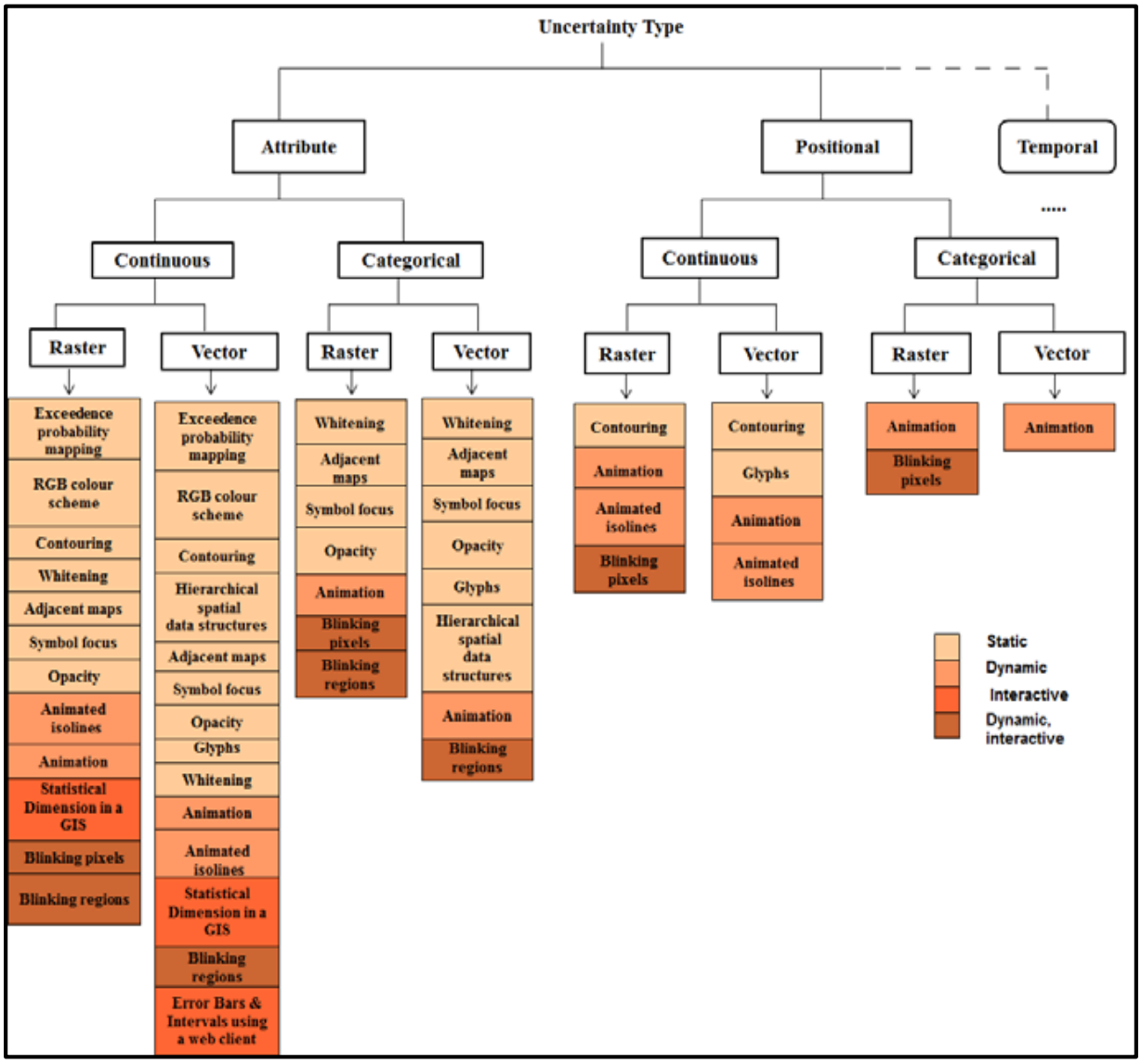
It has been found that the understanding of uncertainty expressed in statistics (accuracy assessments) are still limited to experts and those with strong mathematical backgrounds (Zhang & Goodchild 2002). The use of visual methods to show uncertainty rather than

textual forms, can provide a better and more user friendly communication of uncertainty (Zhang & Goodchild 2002).

Uncertainty visualization is a special form of visualization specifically focussed on the uncertainty in data. It aims at presenting quantified uncertainties in a visual context for understanding (Senaratne et al. 2012). As previously discussed, maps can be perceived as being without error of any sort. However, reality is that uncertainty can arise in a number of different ways and is present in all spatial datasets (Slocum et al. 2013). Five types of uncertainty commonly mentioned are lineage, logical consistency, completeness, positional accuracy and attribute accuracy, with the last two perhaps the most important for uncertainty visualization (Slocum et al. 2013; Hunsaker et al. 2001).

In brief: 1) lineage is the historical record of the digital data, such as who created it, when and for what purpose; 2) logical consistency is similar to topological correctness and concerned with matters such as whether all polygons are closed; 3) completeness refers to whether data such as selection criteria for classes and class membership was indicated; 4) positional accuracy is the concern of the locational accuracy of an object both horizontally and vertically such as a trig beacon; 5) attribute accuracy refers to the accuracy of features found at certain locations, such as if an attribute is cornfields or grassland (Slocum et al. 2013; Thomson et al. 2005).

Senaratne & Gerharz (2011) suggested a model for choosing which method of uncertainty visualization to use, based on the uncertainty and data type that is to be visualized, as seen below in Figure 2.2.



Source: Senaratne & Gerharz 2011

Figure 2.2 Categorization of uncertainty model

Senaratne & Gerharz (2011) first break down uncertainty type into three broad categories of attribute, positional and temporal. Attribute and positional types are further broken down into either continuous or categorical data, and further into visualizations for raster or vector data. Recommendations for visualizations are then made that can be either static, dynamic, interactive or dynamic interactive.

This model can be used as a guide into which form of visualization is appropriate, once the underlying data type and usage choices have been made. Before uncertainty can however be visualized, it must first be quantified through statistics or some other method. The International Bureau of Weights identifies two types of methods for evaluation of uncertainty (Alberti 2013; BIPM 2008): Type A and Type B. Type A, is by the statistical analysis of observations, Type B is evaluation by means other than statistical methods (Alberti 2013).

Alberti (2013) further breaks down Type A and Type B methods. Type A is broken down into 'fuzzy' methods and 'probabilistic' methods. Fuzzy methods are methods such as clustering, similarity selection and Fuzzy Set approaches (Bordoloi, Kao & Shen 2004). Probabilistic methods use analytical parametric or statistical distributions. These include simulation models such as Monte-Carlo and simulation based resampling. Type B is explained as using expert elicitation to analyse uncertainty (Alberti 2013). Type B is more subjective, whereas Type A is based more around objective methods that rely on statistics and scientific exploitation.

Visualization is achieved in a variety of ways. The work of both Alberti (2013) and of Senaratne & Gerharz (2011) highlight different data needs to be treated differently. The type of uncertainty followed by the type of data and the format (vector or raster), dictate the options that are available for the visualization of uncertainty. Therefore, a blanket solution does not exist for visualization, only tailored solutions for data type. Type A methods are better for visualization, as they provide scientifically backed methods either through statistics or simulation that can be expressed for visualization.

Once uncertainty has been determined and a method of visualization has been identified, this visualization must be applied to the data. Two methods of uncertainty visualization currently exist, namely intrinsic and extrinsic.

2.5.1 Intrinsic and extrinsic methods

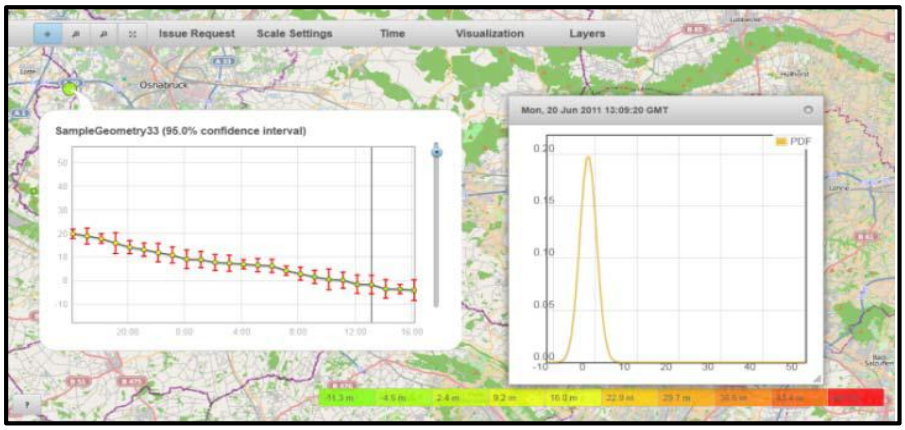
Intrinsic representations form part of the data visualized and can be seen as 'fuzzy edges' or 'varied sizes' of variables and vary an objects appearance (Slocum et al. 2013; Slocum et al. 2003). Extrinsic representation is when additional symbols are added over the dataset, such as the addition of dots or haze over objects (Slocum et al. 2013; MacEachren et al. 2005). Cartographers have been seen mostly to prefer intrinsic visualization over extrinsic forms.

Slocum et al. (2003) found in his study of water balance experts, that experts prefer extrinsic visualizations, whereas decision makers prefer intrinsic methods. It was further noted that

intrinsic methods are best for displaying the 'bigger picture', whereas extrinsic methods were found more effective for indicating the specific locational uncertainty information (Bostrom, Anselin & Farris 2008; Slocum et al. 2003). The choice of which thus depends on the user and the intended purpose of the dataset. Thus, if the aim of the visualization is both to communicate, as well as discover potential relationships between a datasets quality and the spatial component of the dataset such as was done by Lee (2009), using an extrinsic visualization would be more effective. The spatial component of data quality can thus be visualized spatially in complement of the accuracy assessment that gives global statistics.

2.5.2 Methods for visualization

Various tools for uncertainty visualization exist using an assortment of methods similar to those suggested by Senaratne & Gerharz (2011). Some of the more common methods used in tools are adjacent maps, error bars and confidence intervals (Gerharz et al. 2012; Senaratne & Gerharz 2011). Howard & MacEachren (1996) developed R-VIS, one of the earliest uncertainty visualization tools for the specific purpose of uncertainty visualization in a '*kriging*' process. Another tool Aguila, is an interactive statistical tool for uncertainty visualization in raster data (Slocum et al. 2013; Senaratne et al. 2012). R-VIS is however, not available for use to the public, Aguila on the other hand as part of PCRaster freely available to all. UncertWeb is another online tool that is theoretically available however not yet published. Figure 2.3 shows an example of error bars created with UncertWeb.

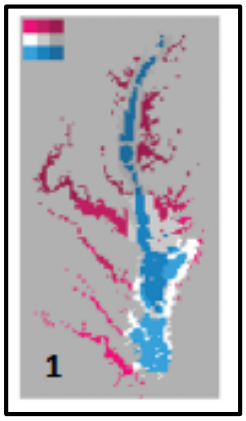


Source: Senaratne et al. 2012

Figure 2.3 Error bars

Error bars are used as a means to show the distribution of values that cause uncertainty. Symbols can also be used to show uncertainty and create awareness of uncertainty (Senaratne et al. 2012; Fowler 2011). Another method used by the Aguila tool is to put an extrinsic

overlay over the data as medium sized dots. These dots grow in size based on their uncertainty value; the higher the uncertainty at the area the bigger and more red the dot. R-VIS by Howard & MacEachren (1996) used a different technique: using the variance estimate for a '*kriging*' process they found that the use of different colours for levels of certainty provided an easier to understand visualization than by using hue and saturation. The result can be seen in Figure 2.4, with it being easy to distinguish between the areas of high and low certainty.



Source: MacEachren et al. 2005

Figure 2.4 R-VIS image

In this visualization, blue is the area of high certainty, with the red denoting areas of lower certainty. It is easy to interpret and quickly provides information about uncertainty to the person viewing the data.

Although many studies have been done on uncertainty visualization, as well as on which methods are most effective, there is yet no single standard or commonly accepted method for visualization (Senaratne et al. 2012). Different tools support different data types and data structures. Each tool has its place within the uncertainty visualization framework. There are working models as well as frameworks for how to visualize a particular piece of data, but a unified tool for all types of data and uncertainty is not yet available. Further, there is yet no *de facto* standard for which tool to use with which type of data.

2.6 VECTOR VS. RASTER

Spatial data is stored in two formats: vector and raster, each with a direct impact on what type of visualization can be used (Senaratne & Gerharz 2011). Described in this part of the study is the difference between the two types of datasets and how each has differing qualities that lead to differences in uncertainty visualization techniques used (Senaratne et al. 2012).

Vector data consists of lines, points and polygons (Congalton 1997). Each vector feature is also associated with a record in the attribute table (ESRI s.a.b; Davis 2001). It is used to represent data with discreet boundaries such as cadastral maps (ESRI s.a.c). Raster data on the other hand, is an array of equally sized cells arranged in rows and columns. Each cell contains an attribute value and location coordinates. These values can be simple binary representing membership of a class or non-membership, or be as complex as a 0 or any other number, when classified such as in a remotely sensed dataset; in this case cells that have the same value have the same geographic feature (ESRI s.a.d; Congalton 1997).

Vector and raster data have advantages and disadvantages. For vector data these advantages are: 1) data can be represented at its original resolution; 2) data can be more aesthetically pleasing and easier to change symbology such as changing from stars to squares for point data: 3) higher geographic locational accuracy is kept through explicit x,y coordinate storage: 4) one is able to efficiently encode topological information and relationships (Buckley 1997). Raster data advantages are: 1) geographic location of cells are implied from control points and the position of the overall cell matrix; 2) data analysis is easy to program due to storage techniques of data; 3) data structures are better suited to quantitative analysis and mathematical modelling: 4) discreet and continuous data are accommodated equally well (Buckley 1997). Disadvantages of vector data are: 1) all vertexes need explicit x,y coordinate storage; 2) topological structures can be complex and processing intensive when doing analysis and as topology is static and any editing of the data requires a rebuild of the topology; 3) continuous data such as rainfall or elevation is not easily represented; 4) it is impossible to filter or do spatial analysis within polygons (Buckley 1997). Raster data disadvantages include: 1) the resolution is determined by the cell size; 2) network linkages are difficult to establish and resolution impacts the efficiency and adequacy of linear feature representation; 3) raster data only defines one attribute per pixel thus attribute data associated per pixel can become complex to store and programme; 4) a large percentage of data are still in vector formats and need conversion which can be problematic as described below; 5) cartographic quality needs are not always applied to maps from the grid-cell system (Buckley 1997).

Although it is possible to convert from vector to raster or vice versa, there are obstacles such as the grid cell size of the raster, the shape of the polygon, shifting of polygon or pixel locations (Lacroix 2009; Congalton 1997), thus creating more uncertainty. It is also important to note, that while it is possible to see raster data grain, vector data looks clear even when it is

zoomed into the dataset (Congalton 1997). Even though vector data looks like it retains its crispness, data was captured at a certain scale and was intended for use only at that scale, thus when changing the scale, more uncertainty and inaccuracy is brought in (Slocum et al. 2013; Kraak & Ormelling 2011; Congalton 1997).

Although both raster and vector data structures have their place in GIS, spatial modelling still prefers the use of raster data (Conolly & Lake 2006). Raster data is often used in environmental modelling especially those that use continuous data such as DEMs or rainfall data that can be used to derive products. The way in which raster data encodes values to each cell without consideration for the theme they represent, further enhances their usability in environmental models, as geometrically indexed vector objects force segmentation of information into different layers whenever they interact in space or time (Yuan 1996). As data structure and display techniques are different, the research will only focus on raster data further, thus there is only a description of visualization for raster data.

2.7 RASTER DATA VISUALIZATION

Raster data can have either intrinsically or extrinsically visualized uncertainty. One method of intrinsically visualizing uncertainty is linked to data classification, using the decision systems that classify the dataset to have specific uncertainty classes. Pixels can be coded as being of uncertain value based on the threshold of the classification system in fuzzy classification systems or interpolated into a smooth continuous form (Shi 2010; Lucieer 2006). This however requires modifying the original input dataset and producing a dataset that may not be immediately usable in a standardised workflow. Another method would be to build uncertainty values into the raster attribute table.

Extrinsically visualizing uncertainty involves adding on top of the dataset. Although cartographers prefer intrinsic visualization, extrinsic visualization has its merits particularly for communicating the spatial nature of uncertainty (Slocum et al. 2013; Bostrom, Anselin & Farris 2008; Slocum et al. 2003). One major advantage is that input data is left as-is and will still be compatible with any models already in use. The uncertainty overlay would thus be a method of communicating information about the dataset as a true visualization, as per the definition of MacEachren (2005), in that it provides information that has to be personally encoded to create new meaning about the original dataset.

2.8 ACCURACY ASSESSMENT AND UNCERTAINTY VISUALIZATION METRICS

Statistics are used when calculating accuracy. Van Niekerk (2016) used three statistics to report the accuracy of the SUDEM: mean absolute error (MAE), standard deviation, and the 90th percentile. Other statistics also explained here are RMSE and z-score (normalized score).

MAE and RMSE are frequently used to evaluate the quality of models (Chai & Draxler 2014; Mashimbye 2013). MAE is the average magnitude of all the measured errors, thus negating for positive and negative values and only using the magnitude of the differences between observed value and reference value. All individual differences are weighted equally and the MAE thus returns the average error of all the errors (Chai & Draxler 2014; Yaffee & McGee 2000). RMSE is the square root of the average of the square of all errors. The square root values are used to accommodate differences that are both negative and positive. Higher values indicate larger error which, due to squaring of the errors, will always affect the resulting RMSE more (Congalton & Green 2008). RMSE gives more weight to larger errors and therefore penalizes larger variances (Chai & Draxler 2014). The RMSE will always be bigger or equal to the MAE; the larger the difference between the two the larger the variance between errors (Willmott & Matsuura 2005).

RMSE is regarded as a standard measurement for model errors (Chai & Draxler 2014; Chai et al. 2013; Savage et al. 2013; McKeen et al. 2005). However, not everyone agrees that RMSE is always useful, suggesting that in some cases a low RMSE score may not indicate good data (Mentaschi et al. 2013). This view is taken, because MAE can be kept at a constant regardless of variance, whilst RMSE may fall or rise due to changes in variance (Chai & Draxler 2014; Willmott & Matsuura 2005). MAE is regarded as a more stable measure of error as it is not affected as much by variance, making it possible, without ambiguity, to compare two MAEs (Chai & Draxler 2014; Willmott, Matsuura & Robeson 2009; Willmott & Matsuura 2005). Some authorities, however, regard RMSE as the better measure of accuracy, as it is more sensitive to extremes than MAE (Mashimbye 2013). In data assimilation applications and calculating model sensitivities, RMSE is still more effective as the penalization of large errors can lead to improvements of models (Chai & Draxler 2014).

Another measure of error is standard deviation. Standard deviation is a statistical measure of the spread of the values from the mean, thus the square root of the variance for a distribution (ESRI s.a.e; Rogerson 2001). It is useful for seeing which values are above or below the

mean score indicating potential patters of under or over estimation. However, a disadvantage is that outliers may skew the mean (Mitchell 1999). Standard deviation also suffers from the same problem as RMSE, in that comparison cannot be between two datasets, but only between one dataset and its verification data. This is because the mean may remain the same but squaring outliers creates a bias towards the outlier (Chai & Draxler 2014; Willmott, Matsuura & Robeson 2009).

Another statistical measurement concerned with the mean is z-score. Z-score is a standardised statistical measure of the spread of values from their mean, in other words a standardized form of the standard deviation (ESRI s.a.f; Rogerson 2001). The standard normal value gives a z-score of zero and a standard deviation from the mean gives a z-score value of one. In normal distributions, 68% of the values would fall within one standard deviation of the mean with 95% having a z-score of + or – 1.96 on a two-sided distribution. The z-score is a measure that can be used to compare different distributions that have different means and different standard deviations (ESRI s.a.f; Rogerson 2001). It is thus useful as a measure to compare datasets, as well as to show outliers. All of the statistics mentioned here however, suffer from a sort of skewing of the mean (Seo 2006; Iglewicz & Hoaglin 1993).

Percentile is the measure of below which value what percentage of values falls. It gives the relative position of a value to the sample. The 90^{th} percentile indicates that 90% of values fall below the value at that point (Frick & Barry 2009). There is a direct correlation between the z-score value and the percentile, thus making the z-score a valuable statistic combined with the percentile to show where uncertainty may occur (Kaplan & Saccuzzo 2008).

Knowing some of the relevant statistics led the study research towards the need for an understanding of how best to represent data visually.

2.9 COLOUR REPRESENTATION

The use of colour can affect data visualization either positively or negatively. This section focuses on how colour is created on paper and on computer screens; it also touches on how colour is understood by humans.

When investigating the representation of uncertainty, colour should be considered as an option along with other options such as texture. What makes this so important, is that visualization is an important method of knowledge creation; throughout the history of science visualization has played a key role in problem solving (DiBiase et al. 2006).

Colour is made up of three components: 1) hue (colour perceived e.g. red, green, blue); 2) value (lightness); 3) saturation (egg vs. sun, amount of pure hue in a colour relative to neutral grey) (Krygier 2014). Of these, hue is often used for qualitative, value for quantitative and saturation for both types of data (Krygier 2014).

There are also various ways in which to see light. One way is simultaneous contrast, such as the way one colour affects the colour next to it (Krygier & Wood 2005). Purity of hues is another way, certain colours look like they are mixtures, whilst others look like they are pure. In this study, a mixture looking colour with varying value could be used to indicate uncertainty (Krygier 2014).

Although humans can perceive millions of different hues, there are further complications when it comes to colour reproduction. Computer screens function on a RGB (Red, Green, Blue) colour system which is additive, whereas most printers function in a subtractive form called CYMK (Cyan, Yellow, Magenta, Black) (Krygier 2014). These factors influence how colour is viewed, making the choice of colours used to represent factors more critical. Another factor is the connotations which humans have with colour e.g., blue = ocean, red/brown = desert, green = vegetation (Conger 2004).

As has been discussed, colour is created in different ways and can be perceived differently. The following section will discuss how a special group of people perceive colour differently to most people.

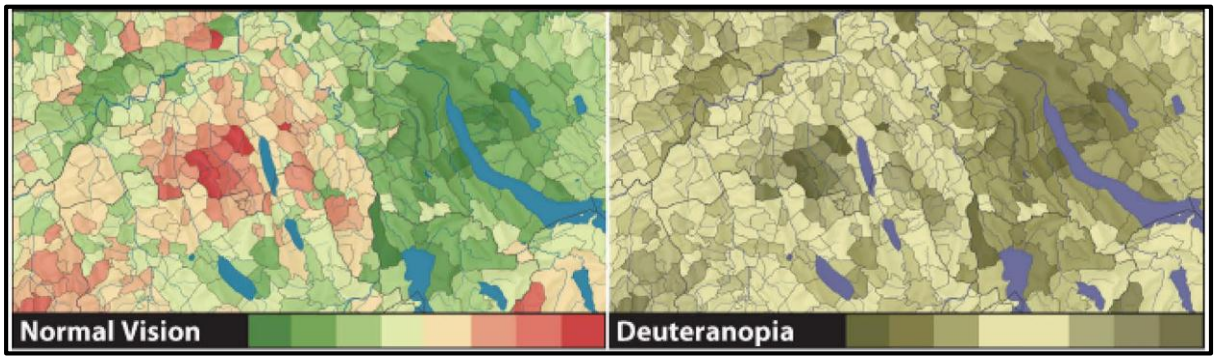
2.10 THE UNCERTAINTY BETWEEN COLOURS: COLOUR BLINDNESS

Colour blindness, clinically known as colour vision deficiency, is more common than most people would expect (Jenny & Kelso 2007). The way in which someone perceives a map also creates uncertainty. With this knowledge, the study focus now turns to how colour blindness affects vision and how this affects perceptions.

Statistics show that one in every twelve males are to be affected by colour blindness (Jenny & Kelso 2007). In Western countries eight percent of males with Caucasian ancestry are said to be affected by colour blindness. Although much less prevalent in females, colour blindness in data visualization cannot be ignored (Jenny & Kelso 2007). A recent multi-ethnic study in the USA has shown that 5.6% of Caucasian boys, 3.1% of Asian boys, 2.6% of Hispanic boys and 1.4% of African-American boys also suffer from some form of colour blindness (American Academy of Ophthalmology 2014).

This is relevant to GIS and cartography as maps are interpreted by the eye in colour. In cartography, colour is often used to create meaning in maps (Brewer et al. 1997). Diverging colours, such as a seven colour red / blue scheme, can be used to show levels above and below a middle point such as grey (Brewer et al. 1997). Such colour schemes are easily interpreted and can quickly provide the reader with key information contained in the map (Brewer et al. 1997). However, as will be presented below, these diverging colour schemes can provide uncertainty in understanding data for people who suffer from colour blindness.

It is important to note, that there are different forms of colour blindness, the two most common being protanopia and deuteranopia (both types of red-green confusion); and even within these two types the intensity of the impairment is large, with some people almost seeing the full spectrum and others having colour blindness in 'pure' as seen in Figure 2.5 (Jenny & Kelso 2007). People who are not colour blind perceive colour through three types of cones, called L, M and S. Each cone type enables the eye to register colour from a different portion of the spectrum (Jenny & Kelso 2007). People whose cones are incomplete, or whose L cones are absent entirely, suffer from protanopia; those with similar problems with the M cone suffer from deuteranopia (Jenny & Kelso 2007). By understanding how colour is perceived, it becomes evident that colour blindness is not a simple condition, but one that varies in type and intensity. For people with the two most common types, protanopia and deuteranopia, shades of purple are well distinguishable; while the other colours blend into shades of yellow and brown. Overall, there is a vast area of overlap where there is no distinction between colours for sufferers of colour blindness. Figure 2.5 reflects what effect this has in practice to visual perception for those with deuteranopia.



Source: Jenny & Kelso 2007

Figure 2.5 Effects of colour blindness

It can be seen from this example, that many colours are perceived as the same, both transition colours and colours at the extremes. Since map reading relies on vision, often together with

colour perception, it is necessary for maps to be produced to enable colour vision impaired users to comprehend the full picture, its meaning and message. Kaye, Hartley & Hemming (2012) indicate that together with making a map visually intuitive, especially for uncertainty visualization, it should be readable by those who are colour blind.

Accepting that colour blindness is a disability, and that people with disabilities are no longer treated as lesser beings, is a symbol of the current era (Olson & Brewer 1997). It is thus essential that map information should be presented to colour blind individuals in a method that they can understand and use with the same ease as the rest of the population (Olson & Brewer 1997). Historically it may have been difficult to produce maps for both full vision and colour blind people. Now, due to the progression of technology, with many maps being viewed on computer screens, it is possible to produce maps that also meet the needs of colour blind individuals (Olson & Brewer 1997). Free software tools, such as Color Oracle, can be used to simulate what someone who is colour blind can see (Jenny & Kelso 2007). Tools such as this can be used to produce maps tailored to the needs of colour blind individuals, as well as maps that are readable by all users, colour blind or with normal vision.

Research such as those by Krygier (2014) and Conger (2004) show that there are many elements to consider when choosing how to represent colour data such as hue, and the way the printer or computer screen produces light. The work of Jenny & Kelso (2007) also highlights that colour blindness and colour perception must be considered as well, which is confirmed by Kaye, Hartley & Hemming (2012); this is especially the case for uncertainty visualization.

Perception is critical when visualizing information; not being able to perceive a dataset in the manner in which it was created can lead to added uncertainty (Kaye, Hartley & Hemming 2012). If uncertainty is to be portrayed in a useful way, then it has to be mindful of all aspects that can cause uncertainty.

In order to compile all this knowledge and the factors or elements involved into one software tool, a structured development plan is needed.

2.11 FRAMEWORKS FOR APPLICATION DEVELOPMENT

Developing an application starts with a design framework. Alberti (2013), in a study developing a web application for the visualization of uncertain spatio-temporal data, identified the use of the data state model specifically for visualization of data. It is based on a backend server side and a frontend client side.

The backend server was described as the UVIS-App (developed by Alberti (2013)), the frontend in the browser as UVIS-Web. As modification of the data state model by Chi (2000), it breaks the work down into four stage operators: the value stage, the analytical stage, the visualization stage and the view stage. The value stage accepts the raw input data; it does not form part of the server side or web application. The analytical stage and visualization stage operators are both part of the server side application. Data is transformed for analytical abstraction based on the analytical stage operators chosen; it is then transformed into a visualization. This transformed data then leaves the server side application and enters the web application, where it is represented visually, so that the end user can toggle the view stage operators to the way the data is represented (Alberti 2013).

The server side application is the core of the whole application. Data is transformed step by step until it is available for viewing on the web by the client. On the client side one finds view operators, which could include functions, such as options for colour blind visualization. The data state reference model breaks down each technique into four data stages and three types of data transformations, with 'within stage' operators also accounted for (Chi 2000). The data state model developed by Chi (2000), and modified in his application by Alberti (2013), was developed specifically for visualization tools; it is the most useful model found for visualization of uncertainty, as it has already been applied in this context by Alberti (2013).

2.12 COMMON GIS SOFTWARE

Two of the most popular GIS packages are ArcMap and QGIS (Friederich 2014; Alberti 2013). QGIS is an open source package that relies heavily on plugins for most of its core functionality. Although the core plugins are developed by the core developers, it also has a very active community creating all types of plugins for additional functionality. There are many resources available for plugin development such as the QGIS application programming interface (API) and books by Westra (2014) and the QGIS Python Programming Cookbook by Lawhead (2015). Plugins can be freely uploaded by anyone onto a central repository and from here can be downloaded by any user (Westra 2014; Lawhead 2015). ArcMap on the other hand is a closed source proprietary software package and has most of its functionality built it. Both packages offer scripting capabilities in Python with ArcMap having a toolbox feature which is similar to the QGIS plugin feature (Friederich 2014). The ArcMap toolbox and QGIS plugins can both be programmed in Python with ArcMap using an enhanced

version called ArcPy and QGIS using PyQGIS. Other open source (Grass, SAGA) and proprietary packages (MapInfo Global Mapper) exist, however these are not as popular as QGIS or ArcMap (Maurya, Ohri & Mishra 2015; Friederich 2014). Thus if a tool for uncertainty visualization were to be developed, it should be for one of these packages. As QGIS is free and open source, it has the potential to reach a larger amount of company's especially smaller consultancies. Maurya, Ohri & Mishra (2015) also found open source GIS software attract better and more productive developers thus support for future developmental tools will be easier to attain. QGIS is however not only linked to smaller companies, local governments in the United States also use QGIS (Repas 2010). The South African government has correspondingly accepted moving towards open source by accepting the Open Source Software policy for Government, thus QGIS effectively is a viable and encouraged choice for an uncertainty visualization tool (Department of Public Service & Administration 2008).

2.13 THE OUTLOOK

This section has taken the reader through the first steps in an introduction to uncertainty as a concept to communicating uncertainty information visually, that would otherwise be viewed statistically. Readers were introduced to academic standards such as the BoK as well as the PLATO model for South Africa, which showed that the South African model is still lacking in both definition as well as scope of what needs to be taught under data quality and uncertainty. Geovisualization was introduced as a way for cartographers and GIS users to show unkowns and new information with datasets, rather than just communicating known information, such as in a 'you are here' map. It was also noted that not all maps fall into either a geovisualization or a communicative category, but that a single map can be either or, or anything in between. Uncertainty visualization was then portrayed as the visualization of uncertain information. Various methods exist, differing from intended use to type of dataset.

Raster data was chosen as the focus for this study, therefore a short introduction was given into how one would interact with raster data during uncertainty visualization. Further colour representation, as well as colour blindness, was looked at in some detail. The way the average human views a colour and how it is produced was examined, as this can cause uncertainty especially when going from the RGB structure of a computer monitor to the CMYK structure of a printed map. Statistics for colour blindness illustrated what people with the most common types of colour blindness experience. It was seen that this is often on a scale from

slightly colour blind to totally impaired. What is important to note here, is that colours in the middle of the colour range are often confused. The Color Oracle software was listed as a possible quick solution to creating maps that are easily understood by people who are colour blind, thus reducing uncertainty. This is important especially in maps that visualize uncertainty, as it would both reduce uncertainty and aid in better visualizing the uncertainty. A brief introduction into application design was given, where the data state reference model was identified as a starting point for the investigation into visualization programming. Finally QGIS and ArcMap were introduced as two of the most popular GIS software, the former being open source and the latter proprietary. QGIS as an open source platform with easy plugin integration and the South African governments push towards open source software was thus identified as a prime candidate for development of an uncertainty visualization tool.

CHAPTER 3 A VIEW OF UNCERTAINTY IN SOUTH AFRICA

From the review of relevant literature, an international view on uncertainty and uncertainty visualization was provided. This chapter serves as the second and final part of Task 1, which is to evaluate the South African perception on data quality. Due to constraints such as time and availability, only a small sample has been selected. This sample consists of those working in the geospatial industry and members of either GISSA or OSGEO's Africa chapter and lastly Stellenbosch University's geography department. Thus it is the view of uncertainty from a select sample of those working in the geospatial industry. As demonstrated in the blue blocks in Figure 3.1 below, this chapter has four subtasks: survey methods, development, conducting of surveys and evaluation of surveys. Together these four subtasks form the second part of Task 1.

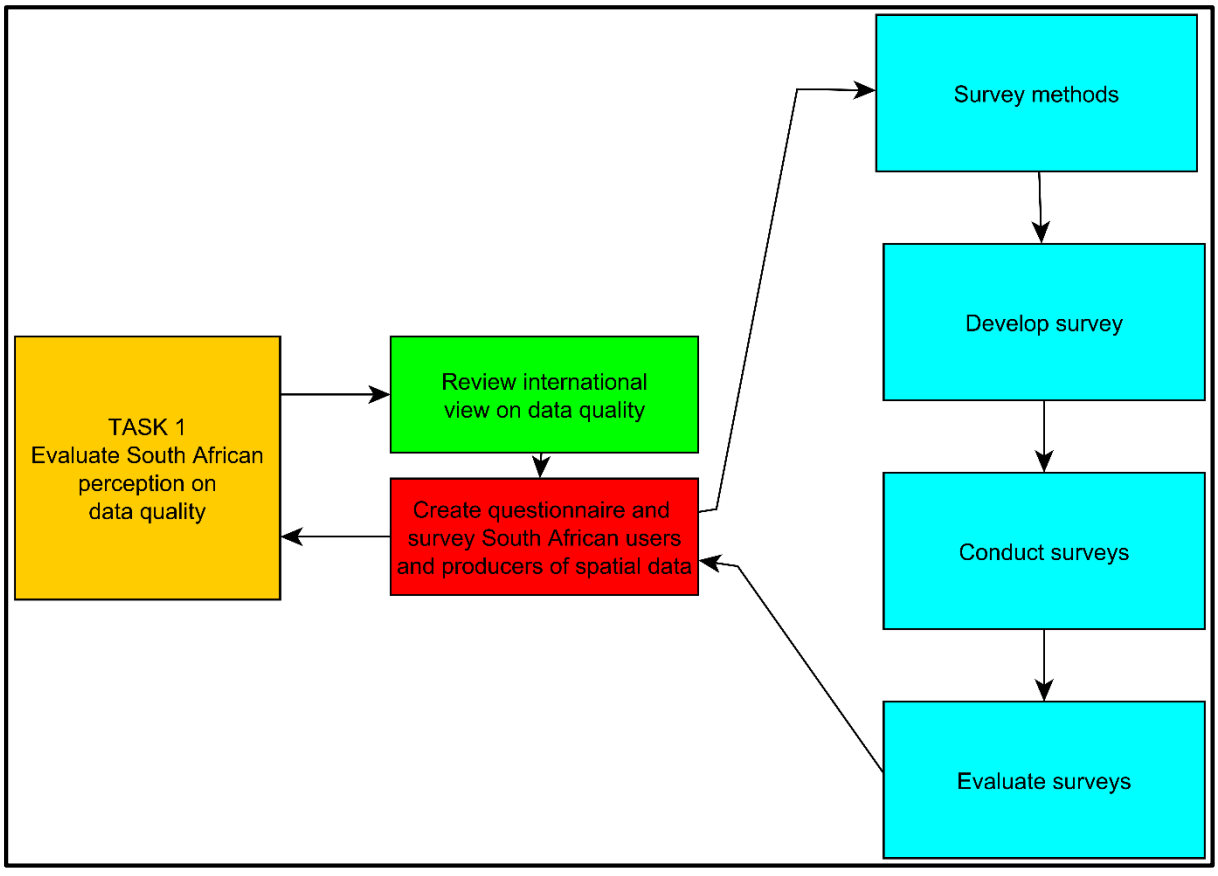


Figure 3.1 Chapter outline

The focus of this chapter is thus to lay the framework for the survey evaluating the South African perception on uncertainty and identify and highlight any limitations of the survey. The chapter ends off with a dissemination of the survey results, to create a picture of what is the view on uncertainty in spatial data in South Africa.

3.1 DEVELOPING THE SURVEY

The purpose of this survey is to evaluate what South African spatial data users and producers understand by data uncertainty and how they manage and deal with uncertainty. For this study users are understood to be those who are using spatial data in geographic information systems (GIS) to derive information, or to create new information from datasets already existing. Producers of spatial data are those who collect and produce raw products which they either pass on to users, or which they further transform into products and then pass on to users.

To gain the insight required for this study, the survey was developed with the following broad questions to be answered:

1) How experienced in years are users and producers?
2) How do they use spatial data?
3) Are they aware of uncertainty in spatial data?
4) How do they manage uncertainty?
5) What form of visualization technique do they find most effective?
6) How do they feel about visualizing uncertainty?

These six questions formed the foundation for developing the more detailed survey questions. They were selected, as together they can tell the story of how well uncertainty is understood and how big a role experience plays in being aware of and managing uncertainty. It was also assessed, whether one group paid more attention and dealt differently with uncertainty than another, as indicated by Knight (1921:289) in Foss and Klein (2012). Question 5 was listed to aid in the development of the uncertainty visualization tool (Uview), so that it may be developed for its target audience and not cause further confusion by being an abstract visualization. Question 6 looked at what the possible concerns of potential users of the tool could be and if they were in line with the predominantly negative sentiment stated by Kinkeldey & Schiewe (2014), as well as De Graaff (2013).

Following this, ethical clearance from the Stellenbosch Research Ethics Committee (REC) was applied for and received (see Appendix A), so that the surveys (see Appendix B) could be distributed. This survey was developed with the following methods and parameters.

3.2 SURVEY METHODS AND PARAMETERS

Research was based on convenience sampling, due to insufficient research time and funding to enable collecting information from a probabilistic random sampling technique (Walford 2011). The indeterminacy of the population size was also a problem, as those working with geospatial data are spread through many industries. In the case of this survey, three calls for respondents were made. The groups chosen were based on all three groups belonging to a geospatial body, so that respondents would have experience and training using GIS and be able to answer the questions as spatial data users or producers.

The first call was made through the Department of Geography and Environmental Studies at Stellenbosch University, the second call through different channels of the Geo-Information Society of South Africa (GISSA) and the third call through the Africa / South Africa Open Source Geospatial Foundation (OSGeo) Local Chapter (a group that focusses on open source geospatial software and news). For this research an online survey method was chosen, some of the benefits entail saving time, there being no interviewer bias or dialects / accents that can affect the respondent's responses that may occur in other methods of surveying and data being automatically collected into a database (Polaris Marketing Research 2012; Nulty 2008). On the negative side however, respondents cannot readily ask for clarification of questions and emails may go astray, whilst self-selection bias can occur (Polaris Marketing Research 2012). Further, online surveys need a motivated group to respond, and only computer literate users are able to respond (Polaris Marketing Research 2012; NEDARC s.a.).

The potential limitation that respondents might not be computer literate is negligible, as geospatial data is by nature related to the use of computers and all users need to have a fair degree of computer literacy. Where it was clear through the responses that respondents misunderstood the question, these responses were used to highlight the issue of uncertainty not being a well understood concept.

The online surveys through combination of the three channels together produced 63 respondents. Although this is a high rate of non-response from the potential of over 2500 respondents from the three channels, which could lead to non-response bias, there is however, no guarantee that it will cause any bias (Baruch & Holtom 2008). Further there is no clear cut-off level that indicates, what non-response rate is too high for using findings in research (Baruch & Holtom 2008; Rogelberg & Stanton 2007). Even a response rate lower than 10% cannot be ignored, especially where there is a previous lack of knowledge, such as the South

African geospatial perception on uncertainty in data (Rogelberg & Stanton 2007). Research such as that of Kinkeldey & Schiewe (2014), which focused specifically on uncertainty visualization, only used a maximum of 12 respondents selected as experts in the field. The 63 respondents are also more than the 44 of De Graaff (2013), however his chosen sample was out of a selected 100 which correlated to a 66% non-response rate. Although the 63 respondents is a high rate of non-response, the works of De Graaff (2013) and Kinkeldey & Schiewe (2014) has shown that the industry is not prone for high response rates, further the findings of Baruch & Holtom (2008) and Rogelberg & Stanton (2007) has shown there is still much value to be extracted from this sample.

3.3 SURVEY RESULTS

Of the total 63 respondents, 19 were female and 44 were male. Respondents were found to use spatial data for a variety of uses. Many of the uses overlapped, thus dividing them into categories was impractical, as it would result in small fragmented groups. The uses of spatial data are data creation, education, analysis and decision making. When evaluating if there was any difference between how difficult the males and females perceived various techniques that were suggested, no significant difference in their understanding of the visuals was found. This was determined by using the averages and comparing them via a t-test: Two-Sample Assuming Unequal Variances with 95% certainty in excel. Therefore the results of this survey can be analysed as one homogenous group.

3.3.1 Sentiment towards uncertainty

General sentiment towards uncertainty was the first piece of information analysed from the 63 respondents. This information was divided into two groups: those with 0-9 years of experience and those with 10 years and more, as both Van Oort (2005) and De Graaff (2013) have indicated that experience plays a role in the understanding of uncertainty. Figure 3.2 below shows that all those with more than 10 years of experience were aware of the inherent uncertainty of data, whereas only 89.5% of those with less than 10 years' experience are aware that uncertainty occurs in all datasets.

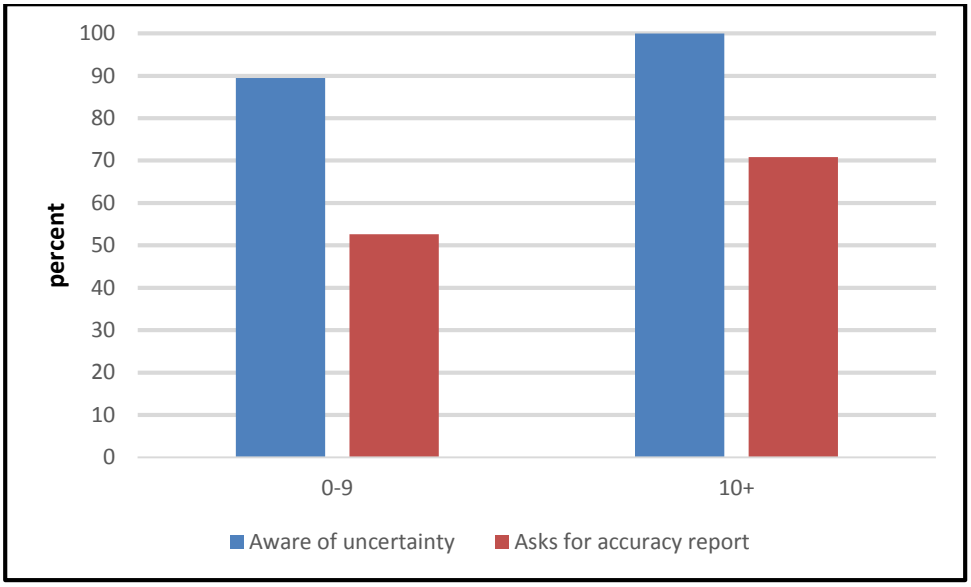


Figure 3.2 Uncertainty awareness for two groups

This supports that there is a good awareness of the presence of uncertainty in data. However, a look at how many users actually seek the accuracy report for a dataset is more concerning. Only 70% of those with more than 10 years' experience request an accuracy report, whereas only some 50% of those with less than 10 years' experience. This could be an issue of apprehension, as GIS is still in the growth phase and universities are annually producing many new graduates into the field annually (McDonough 2008). The concern here is that by not actively seeking to verify the quality of the data being analysed, the resulting data, with its uncertainty, is treated as completely accurate when it is not. The accuracy report / uncertainty element may not even be carried on to users and decision makers down the line. This is especially problematic for building models, as error propagates in potentially a compound consequence (Couclelis 2003). The dangers of neglecting to understand the data one is using was demonstrated in the Netherlands, in a case where a library building was thought to be 1 square kilometre l. This was not just a data quality issue but a lack of knowledge about how the database was created as a value of 1 was the default for unknown size (De Graaff 2013). This is why users and producers of data should always be aware of not only the uncertainty element in data, but also why it was created as well as how it is created and represented.

3.3.2 Frequency of use

When the data was viewed differently by looking at frequent users (daily or weekly users of geospatial data) versus casual users (a few times a month or less), there was little

improvement in the outlook of actively inspecting the quality of data used or created by the study sample: 60% of the frequent users asked for an accuracy report compared to 50% of casual users. Indicating that those who work with geospatial data frequently do not have much bigger concern than those who use geospatial data less often.

Of those who do ask for accuracy reports, only 39.5% would prefer to have a visual representation of uncertainty, whilst slightly more than 70% of those not asking for an accuracy report would like to have a visual method of interpreting uncertainty. Therefore, a well-designed and possibly (in future) standard for representing uncertainty visually, such as the way that Kappa or the confusion matrix has become a cornerstone of statistical accuracy assessment (Foody 2002), can bring the concept of uncertainty to a wider audience. This would vastly improve the understanding of uncertainty and may lead to more people not only being aware of the existence of uncertainty theoretically, but being actively aware of uncertainty within the datasets they use. This is especially true in that most users are aware of uncertainty in a theoretical sense, but almost half do not actively seek to find the accuracy of a piece of data they hold.

A further point of concern is that of those who indicated they develop datasets for use by others. Only 36% ask for an accuracy report on the data they are using to base their work on. This is a problem as to how do they report their accuracy, if they do not give adequate attention to the accuracy of their input datasets. Some respondents however mentioned they work on a ‘fit for purpose’ scheme, which is ideally the way in which spatial data should be used. Fit for purpose in the context of this thesis, is understood as evaluating a dataset to see if it is accurate enough or of sufficient quality for the needs of the model or result required. When working with spatial data, those using it should always consider the scale and accuracy in determining if it is fit for the usage they intend (Aguirre-Gutiérrez el al. 2013; Del Campo 2012).

Skeels et al. (2009) did research on how people from different areas (such as social psychology, computer vision, computer science, bioengineering, radiology, journalism and sales) perceive uncertainty. They highlighted that people conceptualise uncertainty in different ways, although there is some overlap; the phrases ‘imperfect knowledge’, ‘inadequate information’ and ‘lack of absolute knowledge’ reappeared frequently. One participant explained uncertainty as arising from the constant need to fit the processes of the world into models and, when this fit is not perfect, we have uncertainty (Skeels et al. 2009). This is one of the most appropriate descriptions for uncertainty in the geospatial context.

Results from the survey done in this study concur with these research results in that whilst there is a general understanding of uncertainty, different people hold different views of just what is uncertainty. Most respondents used the term as an umbrella term. There was no clear definition given as to what is uncertainty. Uncertainty definition was left open to the interpretation of respondents to find out what were the first thoughts of respondents as to uncertainty in GIS data.

3.3.3 Imperfect data

When respondents were asked how they would feel about knowing that data was 80% accurate (a question deliberately left open so the respondent can apply it to any form of uncertainty relevant to their use), about half responded that this depended on the purpose of the dataset as seen in Figure 3.3 below. Nearly a third of respondents indicated they feel comfortable with 80% accurate data without considering purpose of the data, further slightly less than a fifth of respondents just rejected the data without considering the use. This speaks of the first thought of some respondents not being the purpose of the data, but rather the quality in a vacuum. Low quality data may have application in some cases, similarly high quality data may not be sufficient for very sensitive applications.

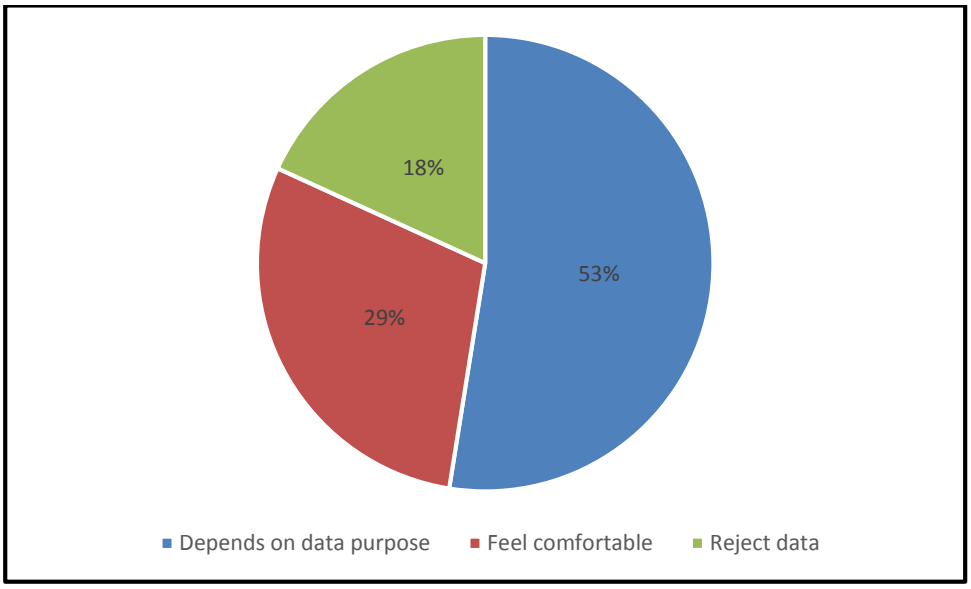


Figure 3.3 How do respondents feel about 80% accuracy?

Figure 3.3 introduces an important result as just over 50% of the respondents even ask for an accuracy report. The longer responses are highlighted further. One of the respondents indicated “In my understanding maps should be 98% accurate because they are mainly used for decision making which will then have an impact on the lives of South African citizens at a

large scale. If one looks at the risks posed by the 20% uncertainty in data we could understand the need for 98%." In contrast to this expectation, this respondent indicated they do not request the accuracy assessment of geospatial data. Further, few datasets comply with this standard - not even the commonly used National Land-Cover of 2000 of South Africa (LC2000), with an accuracy of only 65.8% (Van den Berg et al. 2008). Another respondent indicated "I would be ecstatic, because the raw data that I get to work with is probably not even 60% accurate", which shows a knowledge of the data and accuracy of data that this user is accustomed to. Too low quality is not ideal, however this response also indicated the quality of data often used within South Africa and some assuming to be of high quality. Two other respondents confirmed that they prefer to get the dataset to a 90%+ accuracy. Conversely these respondents indicated that they do not ask for accuracy reports with datasets. About 10% of respondents said they do not trust data with 80% accuracy, but also fall in the group not asking for an accuracy report. This demonstrates a serious lack of understanding of the data that many of these respondents use.

3.3.4 Dealing with uncertainty

How respondents deal with uncertainty proved to be the question with the most varied responses. Table 3.1 shows that most people working with geospatial data try to improve the quality of their data and / or communicate uncertainty in one form or the other.

Table 3.1 How to deal with uncertainty

Response	**Percentage of respondents**	**Percentage of respondents that *does not* look at accuracy assessment**
Try to improve and / or communicate uncertainty	58.7	43.2
Ignore uncertainty and / or guess data quality	19.0	41.7
Fit for purpose	17.5	36.4
Other	4.8	33.3

Of the group that tries to improve and / or communicate uncertainty, 43.2% do not consult the accuracy assessment of the data they are using. The question that remains is how can

uncertainty be communicated or improved upon, when the respondents do not know the quality of the data to begin with. The second largest group of respondents simply ignores uncertainty and / or guesses the quality of the data, even though more than half of this group look at the accuracy assessment, they simply do not sufficiently understand it to deduce if data is fit for purpose. The group that responded they use data on a fit for purpose basis, which should be all users and producers of spatial data, is only 17,5% of respondents, with about a third of these respondents not looking at accuracy assessments is drawn from the questionnaires. Through these varying groups of respondents, it is evident that there are different strategies used to deal with uncertainty, which is in agreement with Knight (1921:289) in Foss and Klein (2012), who states that uncertainty is dealt with differently by different people. Overall, this alludes to the gap in knowledge of the regulations as to how to deal with uncertainty. The lack of a uniform management of uncertainty strategy combined with a large percentage of respondents not looking at the accuracy assessment to begin with speaks volumes as to the inadequacy of the regulations, especially the inability of the local PLATO regulations, at the basic level of expertise required to deal sufficiently with uncertainty. Visualization has been indicated by a large percentage of those who do not look at accuracy assessments as an option they would prefer. A visualization would also be the first representation of the data a user would see. Thus visual communication may thus serve to bring uncertainty to the forefront of thought, as well as communicate it to users and producers from the first moment they deal with data, rather than being hidden away in a separate page in a piece of text known as meta-data, which frequently is lost when new data is generated (Bostrom, Anselin & Farris 2008; Perer & Shneiderman 2009).

3.3.5 Visual communication

If uncertainty is to be communicated visually, then it has to be in an easily understood format. Figure 3.4 represents the question in which respondents were asked to rank how easily they understood these different visualizations of uncertainty. Respondents indicated that Images 2 and 5 were the easiest to understand, even though these two methods are in complete contrast with each other, as seen here in Figure 3.4. Image 5 is a clear representation where a single attribute is shown and uncertainty is represented by different colours, while Image 2 shows uncertainty through interference. Meanwhile, Image 4 was indicated as the most difficult method of uncertainty visualization to comprehend.

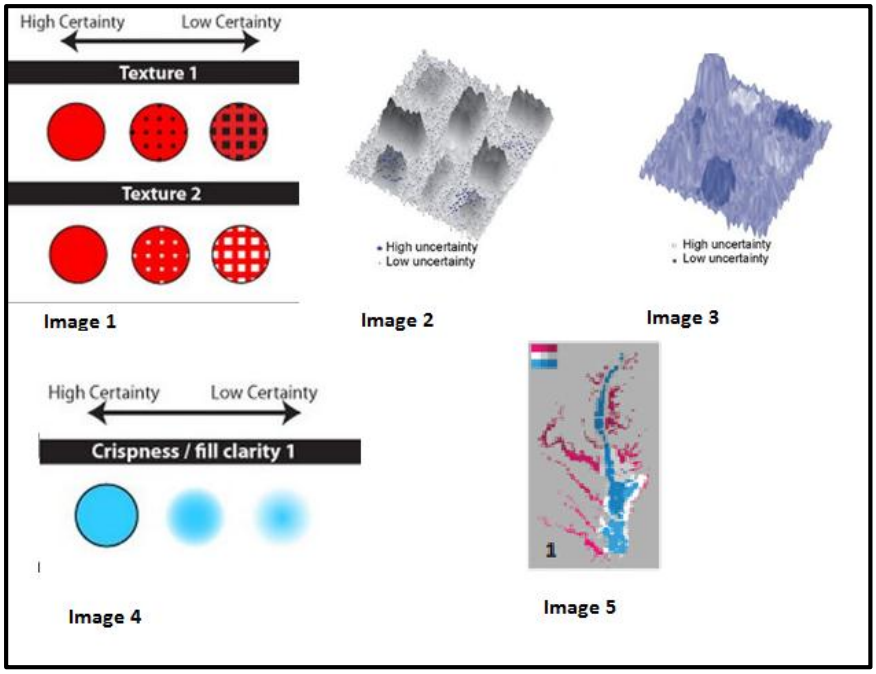


Figure 3.4 Uncertainty visualization techniques

On a positive note, 25% of respondents indicated they visualize data with the colour blind in mind. This indicates that there is a realization for people with colour perception difficulties, even though this realisation is as yet insufficient.

Respondents have indicated that they find the old visualization method from Howard & MacEachren (1996) seen in Figure 3.4, Image 5 to be the easiest to understand. They also indicated that there is some knowledge of visualization for colour blind people, which is to be expected when considering the statistics from Jenny & Kelso (2007), which indicates that one in every twelve males suffers from a form of colour blindness. Comparing the survey findings to the international view on uncertainty in literature completes this chapter.

3.4 FINDINGS COMPARED TO THE INTERNATIONAL VIEW

When looking at the findings from this survey, as well as at international studies, it is evident that none of the studies used a large number of respondents (Kinkeldey & Schiewe 2014; Skeels et al 2009; Tegtmeier et al. 2007). This may be due to difficulties in getting large numbers of respondents, such as time and cost constraints, as well as it being a specialist industry, with busy professionals that often do not have to time to participate in extra research. Rogelberg & Stanton (2007), however, indicated that even surveys with a low response cannot be ignored, especially if there is not much literature in the area.

The results from this survey concur with international literature, that there are no commonly understood meanings for data quality and uncertainty, but rather a hazy notion of what they mean. In terms of the research done in this study, 58.7% of the users and producers indicated they ask for accuracy assessments on data they use. It also became clear that there is a lack of knowledge about uncertainty in datasets, especially by decision makers. Data is often assumed to be of high accuracy by end users and decision makers, without consulting accuracy assessments. This is concerning in the light of the quality of data available, such as the NLC2000 which has an accuracy of 65.8% and the indication by some respondents that they accept high accuracy without considering accuracy assessments (Van den Berg et al. 2008). This agrees with international literature in that there is often a lack of understanding, or just plain ignorance, as to the effect of uncertainty. Most respondents indicated that they know about uncertainty being present in data, which leads to the question of why they do not ask for quality reports. Whilst 70% of respondents who do not request accuracy assessments would like uncertainty to be visualized, only 39.5% of those who do ask for accuracy assessments would like a visualization of uncertainty. These ratios may provide a link with Tegtmeier et al. (2007), who found that some professionals do not want uncertainty visualized as they feel it may reduce the perceived quality of their data, as well as indicate a lack of faith in their own work.

One of the problems with the PLATO model for professional registration as a GIS professional is, that only a small amount of required teaching time (3.36%) is spent on uncertainty, with the result that the presence of uncertainty is widely known but its effects are not well understood (PLATO 2015). There is also no standardised method of dealing with uncertainty other than putting quality information it in a disclaimer or metadata. This may be an answer, but it is not a definitive answer to why most respondents involved with spatial data are aware of uncertainty in principle being inherent in data, but do little or nothing about establishing the true quality of the data, through metadata or other statistical methods. Only about half of the respondents request to see accuracy reports (report about data quality, creation and intended uses) with data, although they nonetheless accept that datasets acquired are of acceptable quality.

Visualization may help in remote sensing tools, where a low accuracy assessment could be achieved, not because of problems in splitting classification classes, but rather the occurrence of miss classification at a particular geographic location. The bivariate method developed by Howard & MacEachren (1996), see Image 5 in Figure 3.4, was indicated by users at all levels

as the easiest to understand, showing that it may be a solid basis for a final representation method. This approach will be developed in the next chapter.

CHAPTER 4 VISUALIZING UNCERTAINTY

This chapter addresses the second and third objectives of Task 2 (see Figure 4.1), namely the review of available uncertainty visualization tools, and the development of a software tool to visualize uncertainty in continuous data.

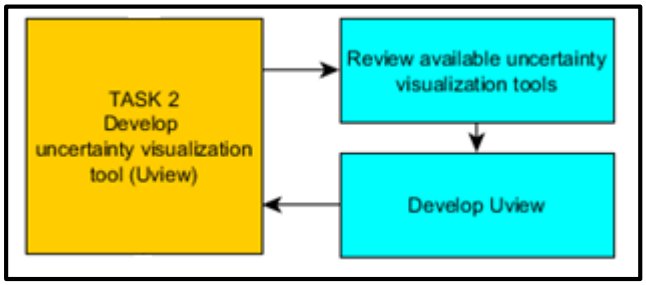


Figure 4.1 Task 2

Before development of the new visualization tool (Uview) could start, the requirements in literature and the capabilities of some existing visualization tools were evaluated (R-VIS, UncertWeb, Aguila and Uvis), along with the Task 1 survey findings in Chapter 3.

Development of the new tool was supported by two key use cases. The first use is for *producers* of spatial data; the second is for *users* of spatial data. Producers of spatial data are here defined as primary creators of spatial data who work with data directly from primary data capture devices such as sensors, satellites and global positioning satellite devices, to produce a dataset. For these producers, the tool creates standard accuracy assessment statistics such as root mean square error (RMSE), mean absolute error (MAE), standard deviation, mean variance and median from the reference points captured for their accuracy assessment (Van Niekerk 2014). The tool further provides an extrinsic visualization of uncertainty based on the calculation of certain uncertainty metrics. This visualization can then be used to either improve the data internally, or be distributed to users downstream with the dataset.

The second use case, users of spatial data are here understood as those, that use already created / modelled spatial datasets to produce derived products, such as a stream delineation from a digital elevation model product (DEM). These users receive a product usually accompanied by statistics explaining the quality of the data, either for their particular study area, or for a larger area, if the product is part of a larger product such as a global DEM. The user would then either collect reference points, or want to compare two products with each other, such as the Advanced Spaceborne Thermal Emission and Reflection Radiometer

(ASTER) DEM and the Shuttle Radar Topography Mission (SRTM) DEM (Huggel et al. 2008). With the tool, the user can then derive statistics, which they can compare with the received dataset statistics, as well as generate a visualization that indicates the spatiality of the uncertainty within the dataset they have acquired.

In conceptualizing the tool referred to as Uview, the South African perspective on uncertainty, as described in Chapter 3, was considered along with international literature. The lack of a suitable and available tool prompted the development of Uview for visualization. This chapter therefore describes the software requirements and selection of the base software environment.

4.1 REQUIREMENTS FOR UVIEW

To create the requirements for Uview, four available tools were reviewed and considered, along with the findings of the survey in Chapter 3. Evaluation of these four tools was based on availability, ease of installation, usability and if it met the needs highlighted in the first survey. These needs can be classified as:

- usability by colour blind individuals;
- availability of a download for further analysis;
- techniques used;
- data requirements for a visualization of uncertainty to be created;
- type of visualization created measured against the choice of the sample in Chapter 3.

The four tools evaluated are R-VIS, UncertWeb, Aguila and UVIS.

R-VIS was specifically developed by Howard & MacEachren (1996) for uncertainty visualization. UncertWeb was a project funded by the European Commission (EC) to create a web client for uncertainty visualization (Gerharz et al. 2012). Aguila is the visualization tool of PCRaster which has been used for uncertainty visualization, particularly by Senaratne et al. (2012). PCRaster is a raster dataset modelling tool that has its basis in the Department of Geography and Environmental Studies, Faculty of Geosciences at Utrecht University; it is open source and available in both Windows and Linux environments (Karssenberg et al. 2010; Pebesma, De Jong & Bierkens 2007). Lastly there is the web based tool Uvis, developed by Alberti (2013) during the course of his research for his master's thesis.

4.1.1 R-VIS

R-VIS, developed by Howard & MacEachren (1996), is one of the oldest visualization tools available. It was developed in 1995 for a specific use case, as a method to evaluate the uncertainty of a nitrogen level dataset. It used *kriging* as its method for developing the visualization. It has been cited by 126 researchers, most recently by Sacha et al. (2016) and McKenzie et al. (2015), thus it can be seen as an effective and still relevant tool. The techniques used are well documented in the paper by Howard & MacEachren (1996); they are still mentioned by MacEachren in recent research (MacEachren et al. 2005). It is, however, not available for download and general use, therefore can only be evaluated based on the information from literature by Howard & MacEachren (1996) and MacEachren (2005). Furthermore, no native support for visualization for colour blind people is supported. Thus, it remains a powerful tool to start off with but, due to unavailability, it is mainly theoretical and not a practical tool that can be used by GIS professionals.

4.1.2 UncertWeb

UncertWeb was part of an EC funded project (Gerharz et al. 2012). It has a solid literature basis and a few proposed methods (UncertWeb s.a.). Being a product of an academic conglomerate project on uncertainty, it had a lot of potential to develop into a largely accepted tool, especially in academia. The project had a set amount of time to achieve its goals. In 2013, the year of the project's planned completion, the project stalled, with a website and the literature still available, but no further indication of progress or any usable tool. A framework developed by the UncertWeb team and described in Bastin el al. (2013) was however published.

4.1.3 Aguila (PCRaster)

Aguila is the primary visualization tool for the PCRaster suite (Karssenberg et al. 2010). PCRaster has been used by Senaratne et al. (2012) for the Aguila tool for uncertainty visualization, but also has many other modelling capabilities; it serves as one of the Faculty of Geosciences at Utrecht University's raster data processing tools, which they continue to improve. It can read many formats of input data and thus can be used with other modelling and GIS software packages (Karssenberg et al. 2010). Aguila can visualize temporal and spatial data, with the added ability of visualizing uncertainty within the data. Thus, it can be used for data analysis and for data exploration (Pebesma, De Jong & Bierkens 2007).

Aguila is a comprehensive tool with advanced statistical analysis and data provided in graphs, with probability, time and cumulative probability options for its visualization (Pebesma, De Jong & Bierkens 2007). While Aguila may be the most comprehensive tool for visualization of uncertainty, it has a few major flaws. Firstly, the installation of PCRaster, as well as Aguila, is no easy task; it requires advanced knowledge and access to administrator rights to install all the dependencies of the software in the Windows environment. Further, the learning curve to use the tool is very steep. The advanced nature of the statistical analysis, as well as an interface that is not very intuitive, may be partly why Aguila is not considered a very popular solution outside of its development institution. It has been mentioned in research about uncertainty and therefore should be evaluated and investigated here (Kinkeldy 2014; Alberti 2013; Senaratne et al. 2012; Gerharz, Pebesma & Hecking 2010). If PCRaster is already part of the GIS user's workflow, Aguila can easily be implemented. However, if the user is new to PCRaster or the stand alone Aguila package, the difficulty of installing Aguila, as well as time lost in development of the skills needed to operate it, may not justify the use of it for uncertainty visualization.

4.1.4 UVIS

Alberti's (2013) work on UVIS was also part of a Master's research project. UVIS has a good academic basis and uses Type A (statistical analysis of observations) probabilistic methods for visualization. It is also a web based tool. However, whilst the tool is incredibly intuitive and user friendly: 1) the visualization scored only moderately in the Chapter 3 survey of this study; 2) Alberti had to be contacted personally to gain access to the tool; 3) the tool was only available as a product demonstration with pre-set data; and 4) no colour blind setting is available natively. It was therefore not possible to use UVIS with one's own data.

4.1.5 Requirements for Uview

The aim is for Uview to be an easy to install tool, especially when compared to Aguila. Uview was therefore developed as a QGIS plugin, because: a) QGIS is an open source GIS package based on the cross-platform library Qt, ensuring that it runs on operating systems such as Linux and Mac OS X as well as Windows; and b) QGIS offers a plugin mechanism which enables individual developers to extend functionality of the main program in a modular way (Shekhar & Xiong 2007). A QGIS user can thus simply install Uview from the QGIS repository with a few clicks (independent of the platform) and easily incorporate it into

their workflow, as suggested by Kinkeldey & Schiewe (2014). To evaluate how Uview would compare, it was evaluated against the four tools evaluated above.

Table 4.1 gives an overview of where Uview is positioned compared to the four software packages evaluated in this study. Only Aguila is freely available, with the others (R-VIS, UVIS and UncertWeb) not having any available implementation to test and incorporate into one's workflow. As Uview will be uploaded to the QGIS repository and be freely accessible for download, it is listed as freely available in this comparison (Table 4.1). Meanwhile, UncertWeb and UVIS were both designed as web applications (WebApps), so theoretically both should have easy access to their functionality as no installation is needed.

Table 4.1 Comparing software

	Uview	**UncertWeb**	**R-VIS**	**Aguila**	**UVIS**
Freely available	X			X	
WebApp		X			X
Easy to install	X	N/A	N/A		X
Easy to use	X	N/A	N/A		X
Provides Statistics	X	X	X	X	X
Colour blind support	X				
Advanced statistical analysis		X	X	X	

From a perspective of ease of use, only UVIS (albeit at a limited testing opportunity) was the most intuitive. Aguila had the steepest learning curve, whereas the skill level required for Uview is no higher than that of the most basic QGIS plugin. All tools rely on statistical analysis to provide a visualization. Further advanced statistics are provided by UncertWeb (according to its literature), R-VIS and Aguila.

Uview provides: 1) the expected statistics for accuracy assessment of a created continuous raster dataset, such as MAE, RMSE and standard deviation; and 2) an easy to understand extrinsic visualization, which does not modify the input dataset and aids in the geographic communication of uncertainty. Thus Uview can easily be incorporated into the workflow of a producer of spatial data, as it provides an accuracy assessment, as well as a visualization, that can be used to communicate the spatiality of uncertainty in the dataset. It can also be utilised by users of spatial data to test datasets quality before data is used, if reference data is available. In contrast to Aguila, no expert knowledge is needed to use Uview, as it only requires basic inputs from the user.

4.2 SOFTWARE TOOL DEVELOPMENT

Development of the Uview uncertainty visualization tool was informed by the initial survey described in Chapter 3, the literature review in Chapter 2 and the comparison of tools inTable 4.1. A QGIS plugin was selected as it is simple to install and use, in line with the suggestion of Kinkeldey et al. (2015) and Kinkeldey & Schiewe (2014) that a tool should be a plugin for QGIS or ArcMAp. These are also the most popular GIS tools, further the QGIS developer community, as open source community was more approachable. Due to the QGIS plugin programming being based in PyQGIS, Python was the language used and enabled easy calculation of the various uncertainty metrics later described. A native option for colour blind people was also a focus, as this would reduce uncertainty in the final visualization (Kaye, Hartley & Hemming 2012).

The Uview tool was therefore developed as a Type A analytical method using probabilistic methods, such as described in Chapter 2. This means Uview: a) uses statistical methods to create a visualization; and b) is used to create an extrinsic visualization of uncertainty which does not change the input data, but which instead gives an extra layer representing the quality of the data over it (Slocum et al. 2013; Fowler 2011).

4.2.1 Framework

The basic development framework, the data state model for application development, proposed by Chi (2000) and also used by Alberti (2013), was followed. The data state model has four stage operators: i) the value stage, ii) the analytical stage, iii) the visualization stage, and iv) the view stage. During the value stage the raw data input is collected; in the analytical stage the raw data is transformed using statistical method; in the visualization stage, data from the analytical stage is used to create a visualization based on the statistics created; and

lastly in the view stage, operators are available methods whereby visualization can be edited and changed to be viewed using different methods (Alberti 2013; Chi 2000). This framework was modified for Uview. Figure 4.2 (below) shows the modified framework used by Uview, with the three main stages of the tool: raw data, data transformation and visualization transformation, with the view stage operator's function integrated into the input phase. Since there is no option to change the visualization from within Uview once it has been run, users of the tool must specify the visualization they would like before running the tool. However, as Uview creates an extrinsic visualization that is delivered in shapefile format containing all the calculated statistics, the user is free to change the visualization as needed and use the statistical field most appropriate outside the Uview tool. In Figure 4.2 Stage 1 represents the raw data input, Stage 2 the data transformation, Stage 3 the visualization.

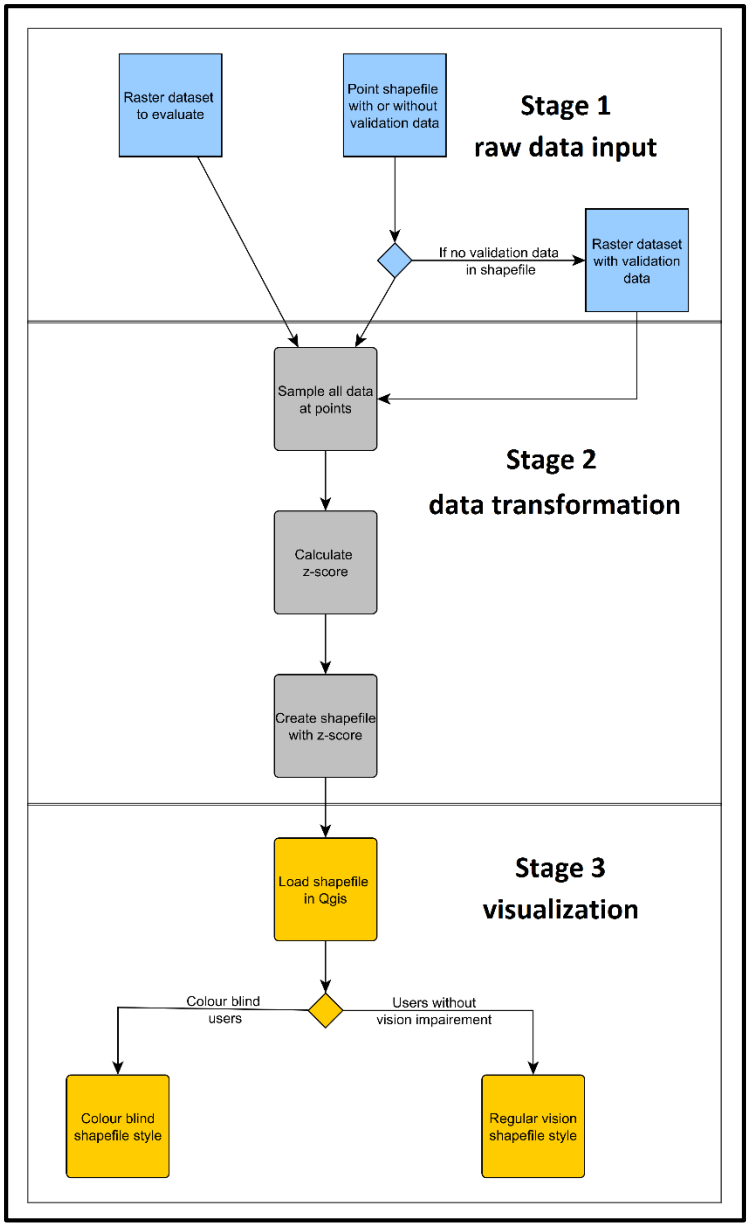


Figure 4.2 Uview framework

All user input is done at Stage 1; this is where the user inputs the continuous raster dataset to be evaluated and supplies a shapefile with points. These points can contain the measured values at these points to be treated as reference values, which will be compared with values sampled from the input raster at these points. If the shapefile only has points, a secondary raster dataset must be supplied; both raster datasets will then be sampled at these points for statistics to be calculated. The user also has the opportunity at Stage 1 to choose whether: a) they would like a colour blind supporting visualization; and b) on which statistical field they would like to have the visualization based. In Stage 2 data sampling is done, and the various uncertainty metrics are calculated and added as fields to a shapefile, then transformed to Voronoi polygons ready for visualization. Stage 3 is where the visualization shapefile is

loaded into QGIS and the data is visualized based on the metric and visualization method selected by the user at the input stage. Whilst this is the standard process for continuous datasets, a process has been developed to make basic provision for discreet data, which will be elaborated on in the discussion on shortcomings.

4.2.2 Development process

As open source software, QGIS comes with a core set of tools, most of which are core integrated plugins; QGIS then relies heavily on external plugins for additional functionality. Plugin development can be done in Python, with Qt (user interface design) used for the user interface bindings. There is also a Plugin Builder plugin to aid in the development of plugins.

To start the development, the PyQGIS application programming interface (API) was reviewed. The first task was to create the Stage 1 interface that would read both raster data as well as vector data. The most useful resources on the PyQGIS consulted throughout the development and coding process were: 1) QGIS API documentation (QGIS s.a.b); 2) GeoApt LLC website (GeoApt LLC s.a.) for Plugin Builder; and 3) the Building Mapping Applications with QGIS book by Westra (2014). Initial development was done in the QGIS Python interface window on test data, as this serves as an instant test bed for code testing.

For Stage 2, the plugin has to load the two or three required datasets into working memory. Before sampling, Uview detects if all required datasets are in the same coordinate reference system (CRS). If they are not, an error message prompts the user to correct this before running the tool again. The next step is to sample the study dataset and the validation dataset, reading these into a Python list and calculating the various statistics. Depending on the input specified by the user, Uview will either: a) read from both study raster and validation raster; or b) read just from the study raster and use the validation values present in the points shapefile. The difference and absolute difference (difference removing the negating negative or positive but just the actual difference from sample to point value) between sample and reference values are calculated. This provides an additional visualization, where it does not matter if the value is higher or lower than the measured value, but only how far away from the reference value it is. Both the differences are added to the shapefile. The standard deviation of the population of the differences is calculated using Equation 4.1.

Equation 4.1 standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Equation 4.1 shows the formula, where σ is the standard deviation, x is each value in the population, μ represents the mean of the population, Σ is the total, and N is the number of values in the population (University of Surrey s.a.).

From the standard deviation, which is a statistical measure of the spread of the data from its mean, the z-score of each point is calculated (Chai & Draxler 2014; Mitchell 1999). The z-score is a standardized measure of a dataset's deviation from its mean. It is normalized so that a z-score of 0 is equal to the mean and a value of 1 is one standard deviation from the mean (Harris & Jarvis 2011). A modified version of the standard deviation is then calculated, where the modified z-score is a statistic similar to the normal z-score. The modified z-score however, uses the median of the difference and the median of the absolute deviation of the difference, to try reduce the effect of outliers on the value (Seo 2006; Iglewicz & Hoaglin 1993).

Each point with its z-score, modified z-score and standard deviation is then written to a shapefile. Other statistics, such as the mean absolute deviation (MAE) and root mean square error (RMSE), are also calculated and written to the point shapefile; both are global statistics used for the measure of accuracy of continuous datasets (Van Niekerk 2014; Hirano, Welch & Lang 2003)

From this point shapefile a Voronoi polygon shapefile is created. Voronoi polygons are areas created around the points, such that any area within that polygon is closer to the point that was at its centre than any other point in the point dataset (Edelsbrunner 2014). This type of inference is based on Tobler's first law of geography that states, "everything is related to everything else, but near things are more related than distant things" (Sui 2003:269). This Voronoi shapefile is then loaded into the QGIS window, and visualized using a style based on the user's choice of style as either a standard visualization or a visualization for colour blind users. By overlaying this shapefile over the input dataset, the user can view which areas contain more uncertainty than other areas.

4.2.3 Uncertainty metrics

Quantifying uncertainty is a way of ranking the quality or accuracy using a percentage scale, rather than with a binary true or false scale. In this way, interpretation of quality is left to the decision makers and their decision-making skills (Burg, Peeters & Lovis 2016). It is a quantitative statement about the probability of error (Foody & Atkinson 2003; Alberti 2013), which must in some meaningful way be quantifiable for visualization purposes (Alberti 2013). For Uview, quantifying of uncertainty for visualization was done using the following metrics: i) absolute difference, ii) the z-score, iii) the z-score of the absolute difference, iv) modified z-score, and v) an overall visualization index (OVI), which will be described in further detail below. Uview as a tool is thus aimed at communicating uncertainty both by providing accuracy assessment statistics, but also by providing a visualization to communicate the quality of the data spatially rather than only globally.

4.2.3.1 Absolute difference

The absolute difference is the simplest and possibly the easiest to understand method of visualization for the end user. It is the absolute difference between the actual value (reference) and the value observed (modelled). It gives a positive difference which can easily be used for visualization. It is however not a statistical measure and cannot be used for comparing different datasets.

4.2.3.2 Z-score based metrics

The z-score is a standardized statistical measure of the spread of values from their mean. It can be used to compare different datasets with different means and standard deviations to each other (ESRI s.a.f; Rogerson 2001). The formula for the z-score is notated in Equation 4.2.

Equation 4.2 z-score

$$Z_i = \frac{x_i - \bar{x}}{sd}$$

In this Equation 4.2 Z_i represents the z-score, while x_i is the observed value and $\bar{x}$ is the population mean, whilst sd represents the standard deviation (Dol & Verhoog 2010; Seo 2006).

Due to it being a measure that normalizes data and allows the comparison of different datasets, the z-score was chosen as the first statistical metric. Two z-score based metrics are

run: the first is the true z-score of the difference (between the reference and test dataset), whether negative or positive; the second is the absolute (difference) values z-score. This absolute value z-score negates the negative values, and considers it as the only true distance from the reference data, so only positive values are used. This affects the mean and standard deviation. As Dol & Verhoog (2010) and Seo (2006) indicated, the z-score may be affected by the effect of outliers as it is based on the mean. This is why the modified z-score, which has been found to be more effective in showing outliers as it uses the median and the median absolute deviation (MAD) instead of the mean and standard deviation, has also been developed as a selectable uncertainty metric (Seo 2006).

The modified z-score is a similar statistic to the normal z-score, as it is also normalized and is comparable across datasets. To prevent the influence of extreme values affecting the z-value, the mean and standard deviation are replaced by the median and the MAD (Seo 2006; Iglewicz & Hoaglin 1993). Before the modified z-score can be calculated, the MAD must first be calculated, by using Equation 4.3.

Equation 4.3 MAD

$$MAD = median(|x_i - \tilde{x}|)$$

In this equation, x_i is the observed value and $\tilde{x}$ is the median and MAD represents the median absolute deviation (Dol & Verhoog 2010; Seo 2006).

The modified z-score for an individual point can then be calculated using Equation 4.4.

Equation 4.4 modified z-score

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

Here x_i is the observed value and $\tilde{x}$ is the median with 0.6745 as a constant with M_i representing the modified z-score (Dol & Verhoog 2010; Seo 2006).

The modified z-score is calculated using the true difference values, as it is the real difference and also accounts for the effect of outliers.

One further metric, named the overall visualization index (OVI), was developed to give a summary view of the z-score based results obtained from Uview. This is the default pre-set value for visualization on the input menu, which is comprised of the absolute difference visualization, the z-score, absolute difference z-score and the modified z-score. The higher the value of OVI, the greater the uncertainty in the input data.

All the z-score based methods were classified into the following categories 0-0.5, 0.51-1, 1.01-2, 2.01-3 and 3.01-. This classification relates to the first class holding 34% of the values, the second class holding 68% of the values, the third class holding 91% of the values and the final classes holding the remaining more extreme outliers. It must be emphasised, that all statistics are always available in the attribute table and the user of the tool is free to visualize on any field and based on any criteria they require. All available visualizations provide guidelines and defaults for a user to perform exploratory spatial data analysis.

An output shapefile is written by Uview which contains the point values of the reference dataset, the value of the raster being evaluated at the point, the actual measured difference as well as the standard z-score, absolute values z-score, the modified z-score, mean, variance, standard deviation, median, MAD, MAE and RMSE. The shapefile also contains an absolute column for the z-score, modified z-score and the difference fields to ease visualization. All visualizations are based on the absolute value fields, as this can be standardized into QGIS style files. The true value remains available for analysis in the attribute table and no data is discarded. This shapefile is then loaded into QGIS.

4.2.3.3 Visualization step

The final transformation step is taking the point shapefile with calculated statistics and creating Voronoi polygons for each point, defining the spatial extent of the uncertainty. Voronoi is the inverse of the better-known Delaunay triangulation (Du & Hwang 1992). These polygons keep the attribute values of the original shapefile and the Voronoi shapefile is then loaded into QGIS with the selected visualization style file applied to it. All the z-score based visualizations use the classification classes described above and shown below in Figure 4.3, with only the absolute difference using Jenks breaks.

Different colour ramps are used for the uncertainty classes, as suggested by the respondents in Chapter 3. Two colour ramps were developed, one for users of the tool with normal vision and one for users who are colour blind. All visualizations have five categories and share the same colour ramp for the two groups respectively. Figure 4.3 illustrates the colour ramp for normal vision as well as the class breaks for the z-score based visualizations.

Figure 4.3 Normal vision colour ramp

Based on the visualization from Howard & MacEachren (1996) seen in Image 5 in Figure 3.4, Chapter 3, this colour ramp was selected by the interviewees as the easiest to understand with the uncertainty classes being distinctly different from one another. The first class is transparent, with only a border around indicating the boundary of the Voronoi polygon. The class that contains the most extreme outliers, i.e. those more than 99.7% away from the mean, is indicated in red. In Figure 4.3, the classes can also be seen for all the z-score based indices (z-score, absolute values z-score, modified z-score and OVI). The values for the absolute difference visualization will be unique to each dataset, as these are classified on Jenks breaks calculated from the absolute difference column.

The visualization style product for colour blind individuals is different, with development based on the literature review as well as the use of the Color Oracle software. What Color Oracle does, is to provide a simulation of what three types of colour blind candidates would see. Only deuteranopia (most common) and protanopia (more common than tritanopia) were simulated. The purple that was selected, was within agreement to the work by Jenny & Kelso (2007). Purple is a colour whose value (lightness) is more distinguishable by people who suffer from these two conditions, as enough distinctions between the values of purple were found for both groups, therefore only one visualization ramp was necessary. As value is also listed as a good way to differentiate qualitative differences, the purple colour scheme was developed. Figure 4.4 demonstrates the colour blind ramp.

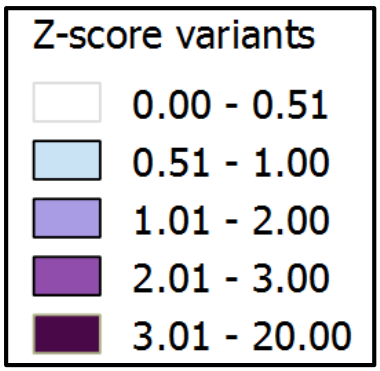


Figure 4.4 Colour blind colour ramp

This colour ramp is found to be effective for both deuteranopia and protanopia. The classes are divided in the same way as the normal vision classes, with the z-score based indices using the classes as illustrated in Figure 4.4, and the absolute difference visualization using the same ramp and five classes, but divided based on Jenks breaks. It was deemed important for Uview to have colour blind support, in order to reducing uncertainty in visualization; also, as Kaye, Hartley & Hemming (2012) indicated, if uncertainty is to be visualized, providing for those who are colour blind is one of the essential requirements.

For the basic discreet data support, only one visualization was created. As it is a binary classification, one colour ramp is used for all users, both colour blind and normal vision. Figure 4.5 shows the two categories, ‘no difference’ and ‘difference’. This classification is not based on statistics as this is an experimental function in Uview for discreet data, Uview was designed primarily for continuous data.

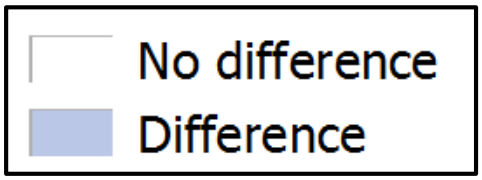


Figure 4.5 Discreet data colour ramp

These visualizations all provide a grouped understanding of the spatial nature of the uncertainty within the data. Uncertainty is put into classes, however as all QGIS functionality is available to the user for the shapefile produced by Uview, the categories can be changed as applicable and users may use any metric. This is the final step in the tool and leaves the user to evaluate their dataset.

4.3 TOOL USEFULNESS

Uncertainty is not a simple concept. The findings from Chapter 3 support the view that uncertainty is not uniformly understood or dealt with in the same manor by all interested parties. Uncertainty is, furthermore, not an easy metric to calculate. Global statistics are still most frequently used for overall accuracy assessments such as RMSE and MAE (Alberti 2013; Mashimbye 2013; Van Niekerk 2014; Hirano, Welch & Lang 2003). Although there is still uncertainty around the exact 'quantifiability' of uncertainty (Burg, Peeters & Lovis 2016; Foody & Atkinson 2003), measures of uncertainty can be treated in a useful manner for visualization. For Uview, three accepted statistics have been used, the z-score, the absolute values of the difference z-score and the modified z-score. In addition, two further visualization options have been included as options, one using the absolute difference between the reference dataset and the dataset being evaluated and the other providing an overall visualization index (OVI) which is the average of the three z-score based indices. Using these five metrics, Uview satisfies its design goals. It is easy to use with a maximum of three datasets required to run the tool. It provides different metrics to enable users to choose particular visualizations. It has visualization styles for both colour blind users and standard users. In addition, Uview generates a shapefile containing global and local statistics for individual polygons which can further be visualized as the individual user sees fit. The ease of QGIS plugin installations from the repository make Uview the easiest to install of all evaluated tools for uncertainty visualization. Thus it can be easily incorporated into the workflow of QGIS users, as suggested by Kinkeldey & Schiewe (2014).

Uview provides producers of spatial data with a method to evaluate and communicate uncertainty in their created data, by using the accuracy assessment capability of Uview, the extrinsic visualization and styles, along with generated statistics to provide better metadata. Users of spatial data can evaluate quality from the global statistics and compare this with the metadata provided by the data supplier. All users of Uview, however get an extra uncertainty visualization capability, used to communicate the spatial aspect of uncertainty. Users can evaluate the quality of the overall dataset, as well as locally at polygon level. This can be used by producers to understand where the problems lie within the dataset, improve the dataset if need be and for users of data to determine if a dataset is fit for purpose. Visualization can also be passed on to users of the data so that they may better understand the quality of the data. As the survey in Chapter 3 indicated, there is a large portion of the users of spatial data who do not understand the quality of their data, and who may even ignore it.

The use of Uview provides an opportunity for all users of spatial data to understand their data better.

CHAPTER 5 UVIEW CASE STUDY

This chapter deals with the evaluation of Uview, the tool developed and described in Chapter 4. It therefore addresses Task 3 as laid out in Chapter 1 and contributes to the fourth objective: to generate visualization scenarios to test the developed software tool (Uview) with. In Task 3, (see Figure 5.1), the focus is on evaluating Uview.

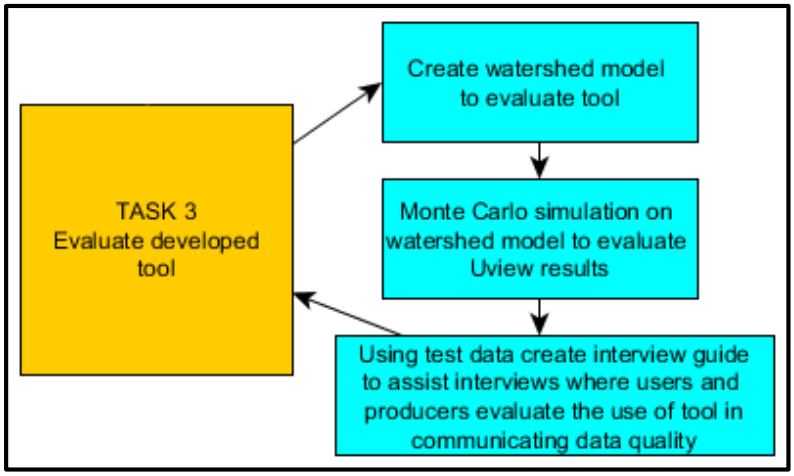


Figure 5.1 Task 3

This is achieved in three sub-sections; the first is to work with the study areas and develop the evaluation data; and the second sub-section of Task 3 focusses on modelling uncertainty in watershed using a Monte Carlo simulation of the spatially autocorrelated error in the Stellenbosch University digital elevation model (SUDEM) (Zandbergen 2011), and is described in detail later in this chapter. This model was compared with the visualization from Uview to evaluate both the effect of uncertainty, in this case of the digital elevation model (DEM) on the generated products, as well as to evaluate the efficacy of Uview as a tool. The third sub-section, qualitative evaluation of Uview (Objective 5), is addressed in Chapter 6.

This chapter is a case study of the use and usefulness of Uview, by firstly introducing Uview as a tool, then by introducing the development of data that was evaluated using Uview. An in-depth analysis of Uview and uncertainty in the DEMs and their product (watershed models) is then discussed with the help of a Monte Carlo simulation based on the work of Zandbergen (2011) and on DEM editing.

This chapter firstly provides an overview of Uview, its installation and general guidelines on using the tool to visualize uncertainty. This is followed by a description of the data to be modelled and naming conventions used for the modelled data. Lastly the power of visualization in conjunction with statistics is evaluated.

5.1 UVIEW INSTALLATION AND USE

Uview is a simple to install QGIS plugin. Once the user has copied the tool folder to the QGIS Python plugin folder, or simply downloaded it from the QGIS repository (once it has been uploaded), Uview is simply installed by clicking 'install' within the Plugin Manager and Install Plugin window inside QGIS. The tool can then be found in either the Plugins toolbar as a graphic in the main window, or on the Plugins dropdown.

The main window as seen in Figure 5.2 shows all the input options.

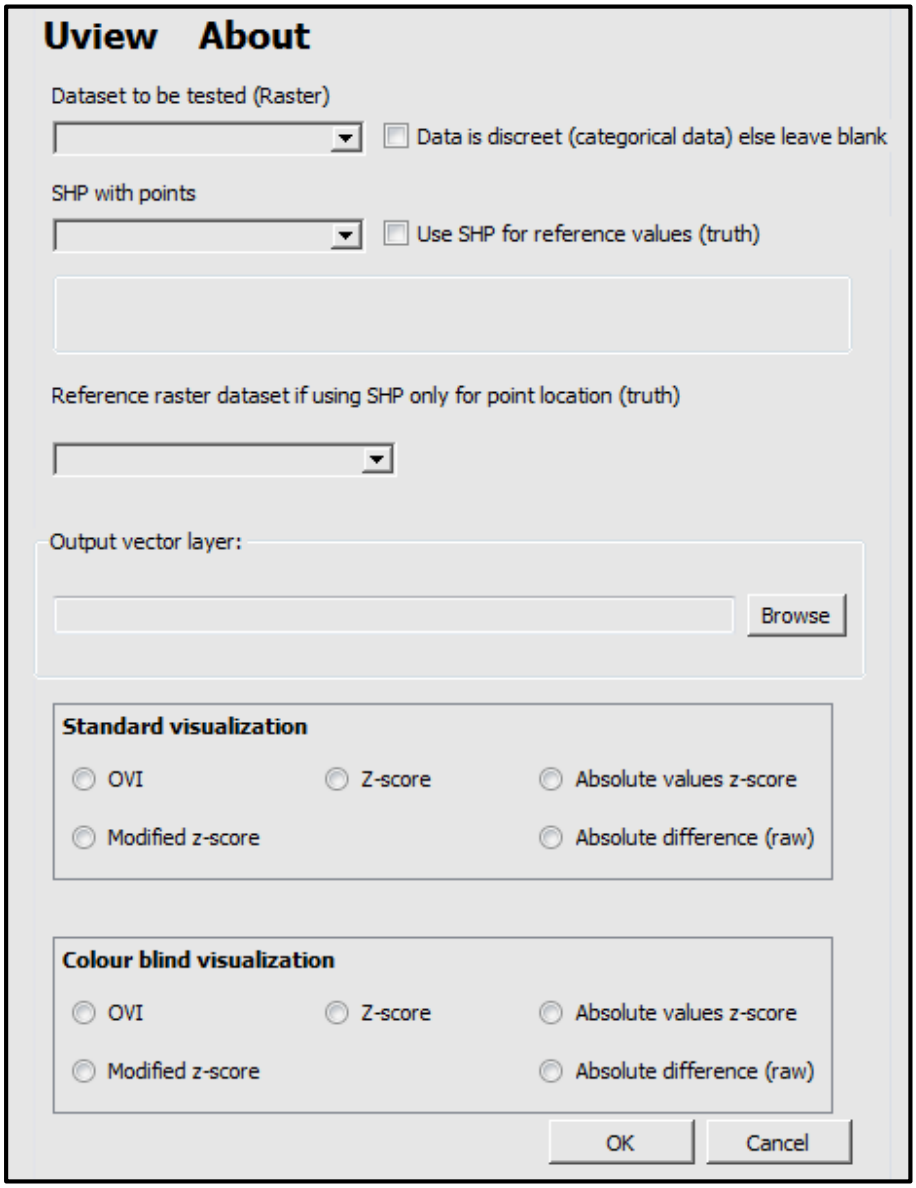


Figure 5.2 Uview main page

In this input screen menu (Figure 5.2), the user is presented with a request for the raster to be evaluated and for the characteristics of the input raster via a check box. This box should be checked (ticked) if the data is discreet and not continuous, as Uview expects continuous data.

The check box will provide a warning (see Figure 5.3) indicating that the tool was specifically designed for continuous data and has very limited functionality for discreet data.

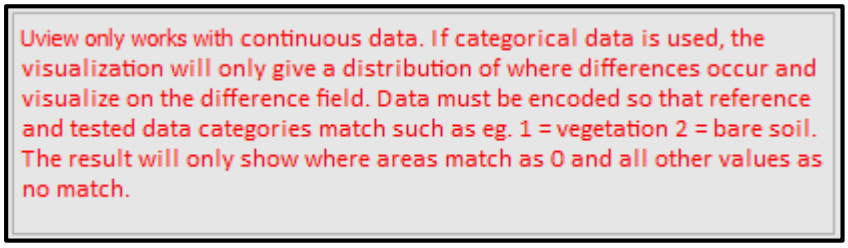


Figure 5.3 Uview main page with box ticked

This basic product for discreet data will create a shapefile only to check whether or not a measurement is correct at a given point. The statistics calculated will be irrelevant (they do not bear meaning and are only generated due to the Uview standard processes) to the end result, as they do not bear relation to the actual value of the input datasets. In this case, visualization is based on the difference between the two values. This is as it is expected: if the two values are the same there would be zero difference, indicating a correct classification at that point. Any value that is not zero is expected to be incorrect and will be flagged as incorrect with no indication of level of uncertainty, as is the case in the continuous data product of Uview.

The second input box in Figure 5.2 requests for a shapefile with points. The check box allows the user to specify if the reference values are contained in this shapefile or not. If 'Use SHP for reference values' is selected, a second drop-down with the fields in the shapefile appears, from which the user must choose the field with the reference values. If the values are not contained within the point shapefile, an input for a secondary raster is provided, which will be treated as the reference raster. The user must also specify the location where the created output product should be stored.

The user is also presented with two blocks with five options each, where the uncertainty metric (OVI, z-score, absolute values z-score, modified z-score, absolute difference (raw)) is chosen for visualization. These two blocks differ only in that the former is for those individuals with no colour vision impairment, and the latter for those who suffer from colour blindness.

5.2 UVIEW REPRESENTATION

Once a user has correctly entered all the required data and clicked OK, Uview runs in the background and provides a visualization in the main QGIS window and visualization key in

the layers panel (LP). Figure 5.4 (below) shows what the user will see once Uview has run successfully.

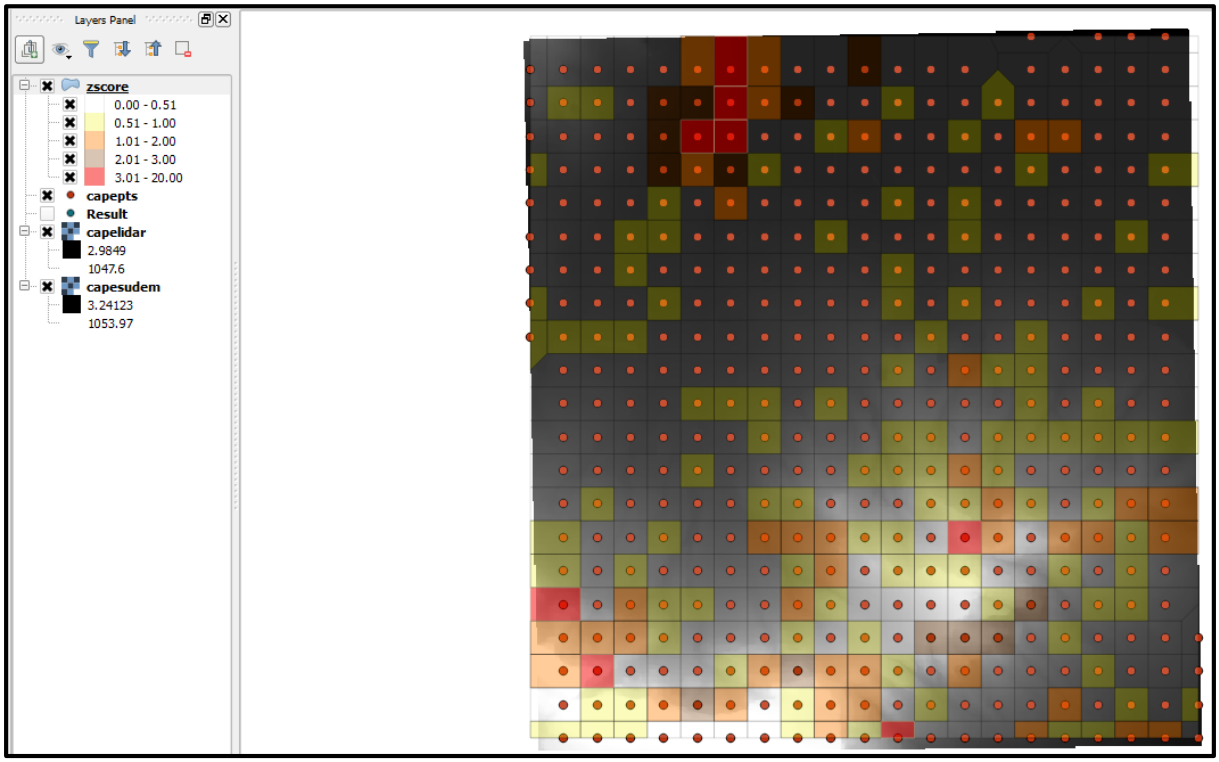


Figure 5.4 Uview product

Based on the colour ramp, the orange, red and brown areas highlight areas of higher uncertainty, whilst the areas that are yellow or transparent show less uncertainty. The LP shows the categories into which the values of the metric have been classified. These are based on z-score values, with values greater than two being outliers. Further investigation into the actual values of these polygons may then be performed. The user may want to use the QGIS identify tool to find all the attribute values at selected points. The user can also choose to open the attribute table and inspect individual records. Additionally, a user may alter the Style used to symbolize the polygon shapefile as demonstrated in Figure 5.5.

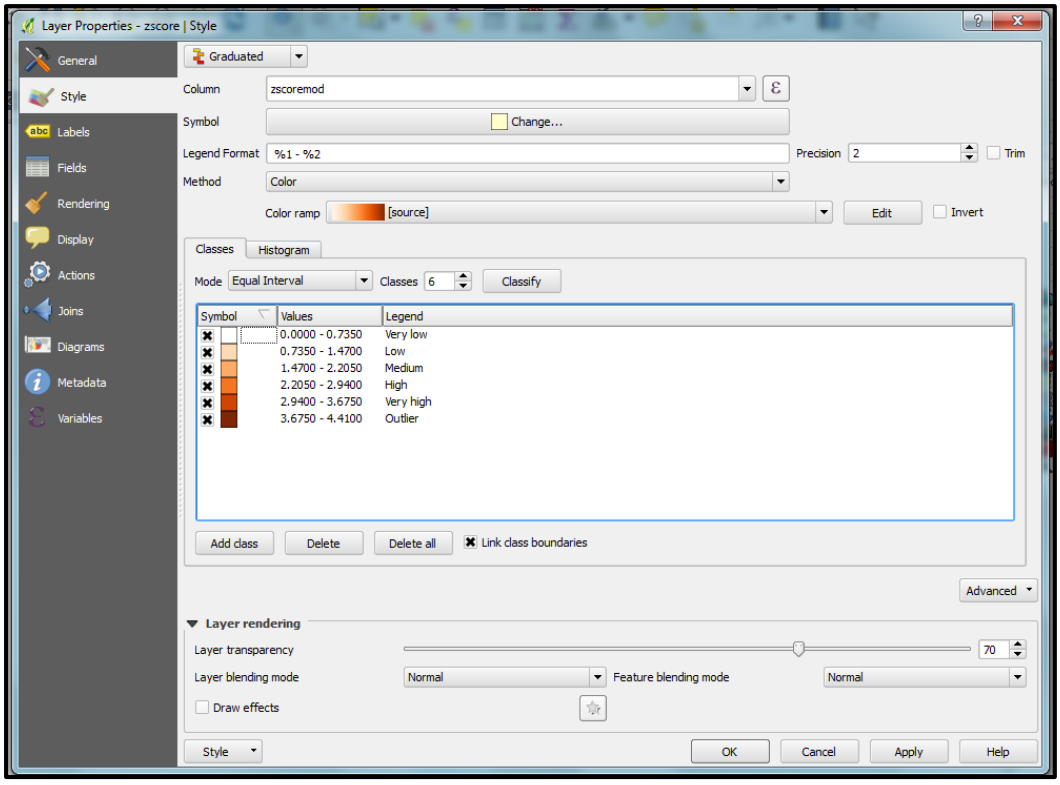


Figure 5.5 Polygon properties dialog

The number of categories, the classification method (Mode) or even the uncertainty metric (Column) to be visualized, can be altered at the user's discretion. If a user selects colour blind visualization, they will be provided with a visualization as seen in Figure 5.6.

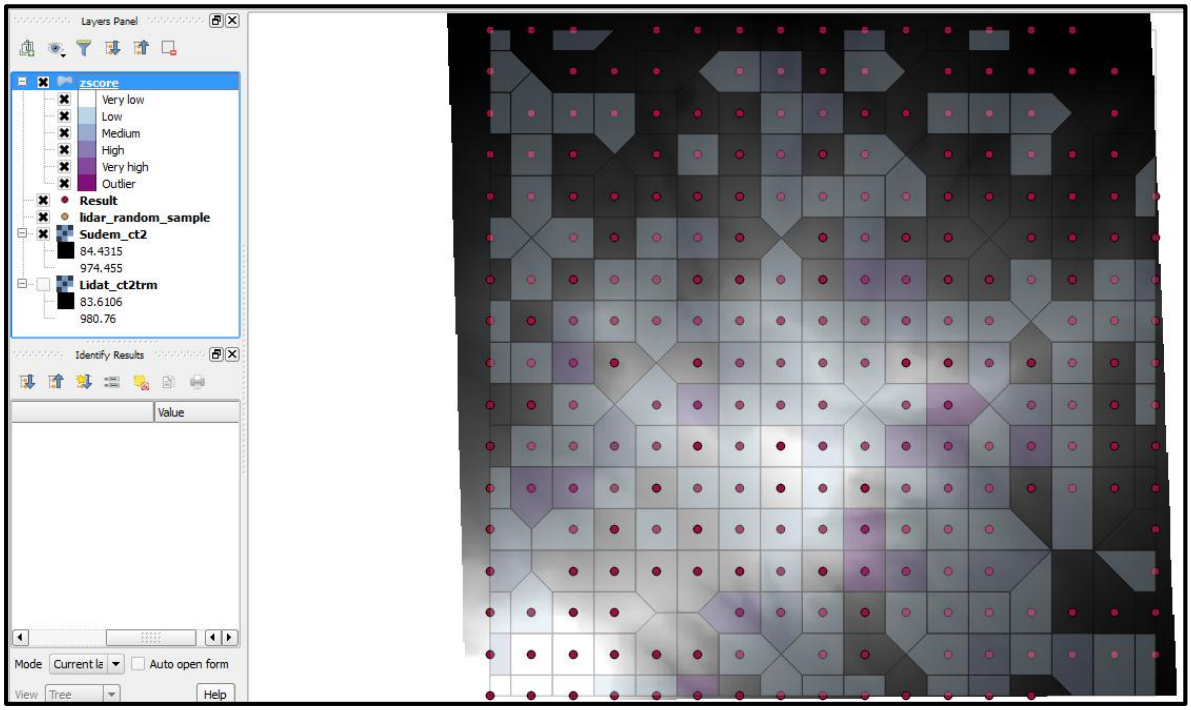


Figure 5.6 Uview colour blind product

This product is sensitive to those with colour blindness as described in Jenny& Kelso (2007). Through the use of the simulation tool, *Color Oracle* it has been found suitable for most colour blind individuals.

Uview has not been developed to cater for discrete data uncertainty, but provides a rudimentary visualization (Figure 5.7) as a rough starting point to determine the spatial nature of the uncertainty. This shortcoming will be addressed in future updates to Uview.

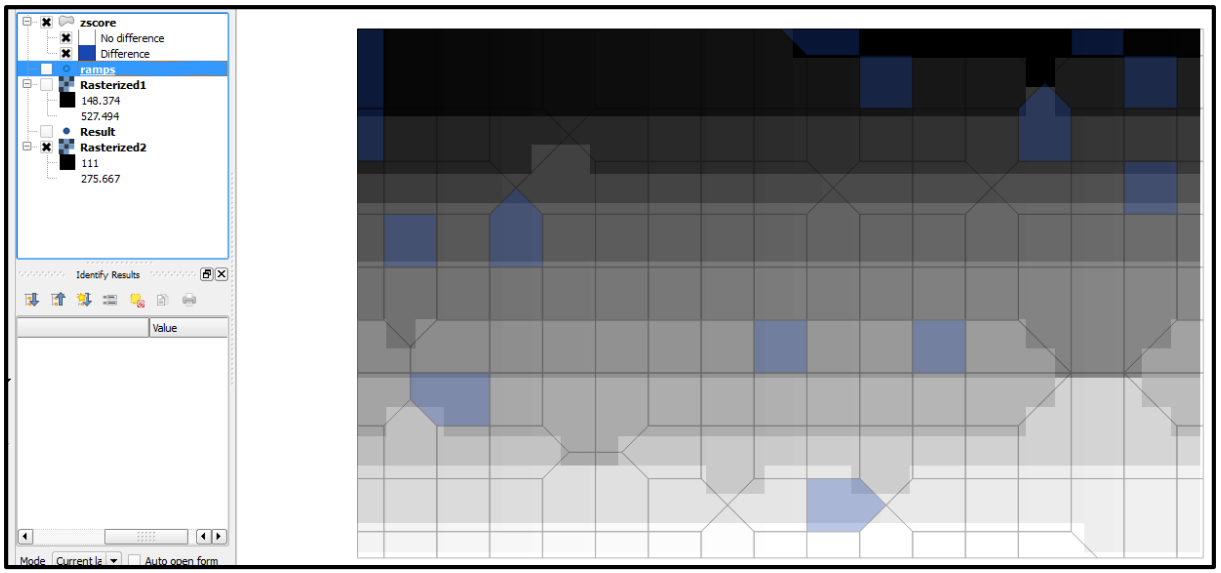


Figure 5.7 Uview discrete data product

Figure 5.7 only shows two areas: those that are correct against the reference data ('No difference') and those that are not correct ('Difference'), thus creating only two classes in the LP. Polygons may even cover the boundaries of two areas. No indication of degree of uncertainty or statistical distance from correct classification is provided either, but the tool may still be useful and is a basis for further research. Further extension to Uview may provide a feature for a confusion matrix to be created from data. As the discreet data product stands currently, both colour blind and standard visualization provide the same style of the visualization, since there was enough distinguishing characteristics between no colour for 'No difference' and the blue for 'Difference' in the *Color Oracle* simulation.

5.3 DATA FOR MODELLING

For this project two study areas were chosen, one on the Cape Town (sea shore to Table Mountain) region and another in the Helderberg region of the Western Cape. Figure 1.1 in Chapter 1 can be consulted for a clear indication of where these areas are. For each study area, two DEMs were compared (test and reference). All DEMs were resampled to 5 m

resolution, the coarser resolution of the two DEMs, to ensure uniform comparative products. In each case the test dataset was extracted from the SUDEM, while the reference DEM dataset was derived from light detection and ranging (LiDAR) data. For more information on the two datasets consult the study area section in Chapter 1.

5.3.1 Cape Town study area

In the Cape Town region, the elevation ranges from sea level up to the top of Table Mountain, with a total elevation change of 1077.7 m. With this steep change in elevation, a difference between the two datasets (test and reference) was expected. The dataset to be tested, the SUDEM (from here on referred to as Test A) has a root mean square error (RMSE) of 3.64 m, a mean absolute error (MAE) of 2.44 m with the 90^{th} percentile at 4.11 m, compared to the LiDAR dataset. This represents the global overall quality of the DEM, which is not spatially explicit. This is important to note, as most statistics are global where a single number represents the data quality with no spatial representation. A visualization is one of the ways in which statistics can be presented spatially. Test A was measured against a LiDAR DEM, which will be referred to as Reference A (Ref A). A LiDAR dataset was chosen as reference as LiDAR has been found to be more accurate than other methods of elevation information acquisition (FUGRO s.a.; Habib &Van Rens 2008).

5.3.2 Helderberg study area

In the second test area, an area in Helderberg in the Western Cape, the SUDEM (in this case Test B) and a LiDAR dataset (Ref B) were again chosen. It is another area where the elevation changes rapidly over a short distance. From the low lying areas up to the top of the surrounding mountains, the total elevation change is 855.6 m, the MAE 2.92 m, RMSE 4.59 m and a 90^{th} percentile of 5.05 m was calculated based on the Ref B.

Table 5.1 provides an easy to access reference point for the naming conventions used for the study areas, as they will be discussed extensively later in this chapter.

Table 5.1 Study area naming convention

Study Area	Name
Cape Town SUDEM	Test A
Cape Town LiDAR	Ref A
Helderberg SUDEM	Test B
Helderberg LiDAR	Ref B

A watershed model was run for each of these datasets and compared using the Uview visualization to demonstrate the power of visualization.

5.4 GENERATE VISUALIZATION SCENARIOS TO TEST UVIEW

This section describes the steps taken to test the effectivity of using uncertainty visualization methods for the two study areas. Uview was run on the DEMs, using Ref A to validate Test A and Ref B to validate Test B with different visualization options (methods). The watershed delineation for the DEMs (Ref A, Test A, Ref B, Test B) is described in detail below. Areas of higher uncertainty based on the visualizations are statistically compared and suggestions are made for selection of the most appropriate visualization.

5.4.1 Watershed from DEMs

ArcMap was used for the watershed delineations, as QGIS proved unstable (crashed on watershed delineation tools) and the Zandbergen (2011) model that will be described and used in the following sections is already tailored for ArcMap. The TUFTS University (2012) model for watershed delineation was followed, as it uses the Basin tool and not the Watershed tool. According to ESRI (2016), the only difference between these two tools is that Watershed requires pour points for delineation, whereas the Basin tool does not. As the pour points for the study areas are not known, the Basin tool was used. Figure 5.8, shows the ArcMap model to create the basins.

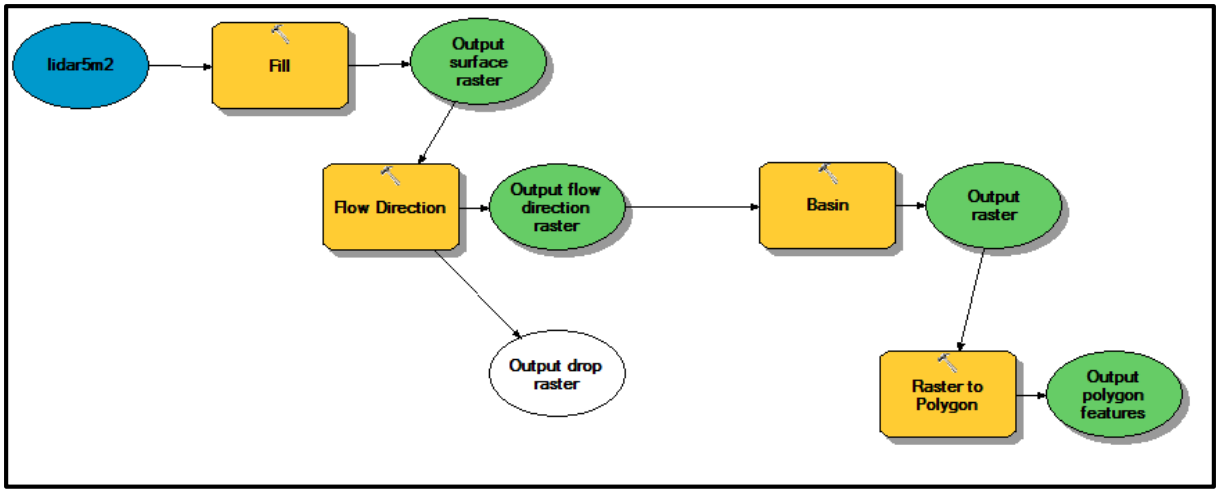


Figure 5.8 Watershed delineation model

The steps for the basin delineation are simple. The sinks in the DEMs were filled and a flow direction was run, followed by the ArcMap Basin tool. Finally, a raster to polygon completed the processing. An ArcMap model was used to ensure processing of all data for the two study areas were the same. Table 5.2 shows how the products for these models were labelled.

Table 5.2 Model data naming conventions

Input Data	**Model Name**
Ref A	Basin-RA
Test A	Basin-TA
Ref B	Basin-RB
Test B	Basin-TB

As with the input data for visualization (Table 5.2 provides easy to access reference point for the naming conventions used for the basin (watershed) products as they will be discussed extensively later in this chapter), Basin-RA serves as the validation data for Basin-TA and Basin-RB as the validation dataset for Basin-TB.

5.4.2 Comparing scenarios

The uncertainty overlays of Uview Test A and Test B were compared to evaluate if and how they relate to the differences between the output Basin-RA – Basin-TA and Basin-RB –

Basin-TB respectively. This was done through comparison of Uview visualizations with histograms, box plots and scatter plots.

Five different methods of visualization were compared: i) the overall visualization index (OVI), ii) z-score, iii) absolute difference based z-score, iv) modified z-score, and v) absolute difference (raw) visualizations. For ease of visualization, the absolute values of all these statistics were used. The absolute difference (raw) was visualized based on Jenks (natural) breaks, which is a standard method for dividing a dataset into a certain number of homogenous classes (North 2009). The Jenks breaks method was developed to minimize within class variance and maximize between class variances (Jiang 2012). If data is skewed to either end, Jenks breaks may create classes with large ranges next to classes with small ranges giving a false impression of data distribution (Shin, Cambell & Burkhart 2016). The z-score based methods were classified according to standard deviations (SDs) into the following categories 0-0.5, 0.51-1, 1.01-2, 2.01-3 and 3.01-. This relates to a normal distribution of the data in which the first class will hold 34% of the values, the second class 68% of the values (up to one standard deviation away from the mean), with the third class containing 91% of the values and the remaining more extreme outliers found in the final classes.

5.4.2.1 Test A

Test A was the first area evaluated. Figure 5.9 shows the absolute difference uncertainty visualization for Test A, which introduces the reader to the first visualization from Uview that gives an introductory view of the data quality. This visualization is based on the raw difference values, so users can gain a quick overview of the maximum true error values.

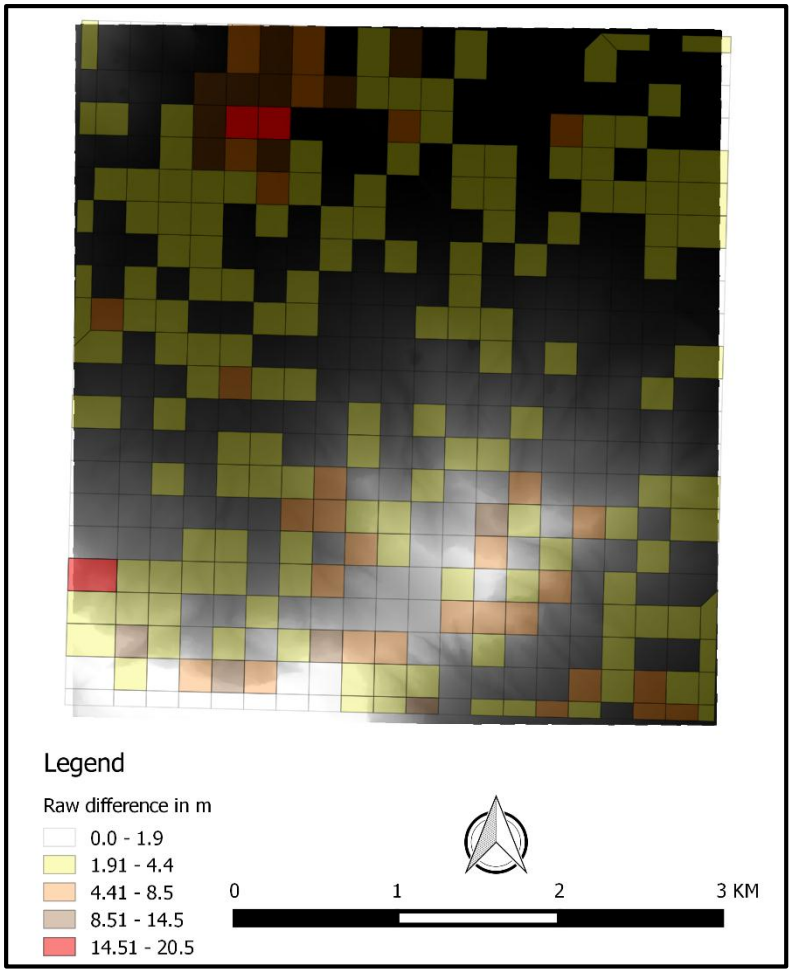


Figure 5.9 Test A overall uncertainty visualization

It is immediately visible that the maximum error is 20.5 m, representing the biggest difference between the two datasets. The visualization does not indicate if this difference is an over-prediction or an under-prediction compared to the reference. The user will have to inspect the attribute table to find this information.

Figure 5.10 depicts all the z-value based metrics. The comparison in this visualization is for the watersheds generated from Test A and Ref A, called Basin-TA and Basin-RA respectively.

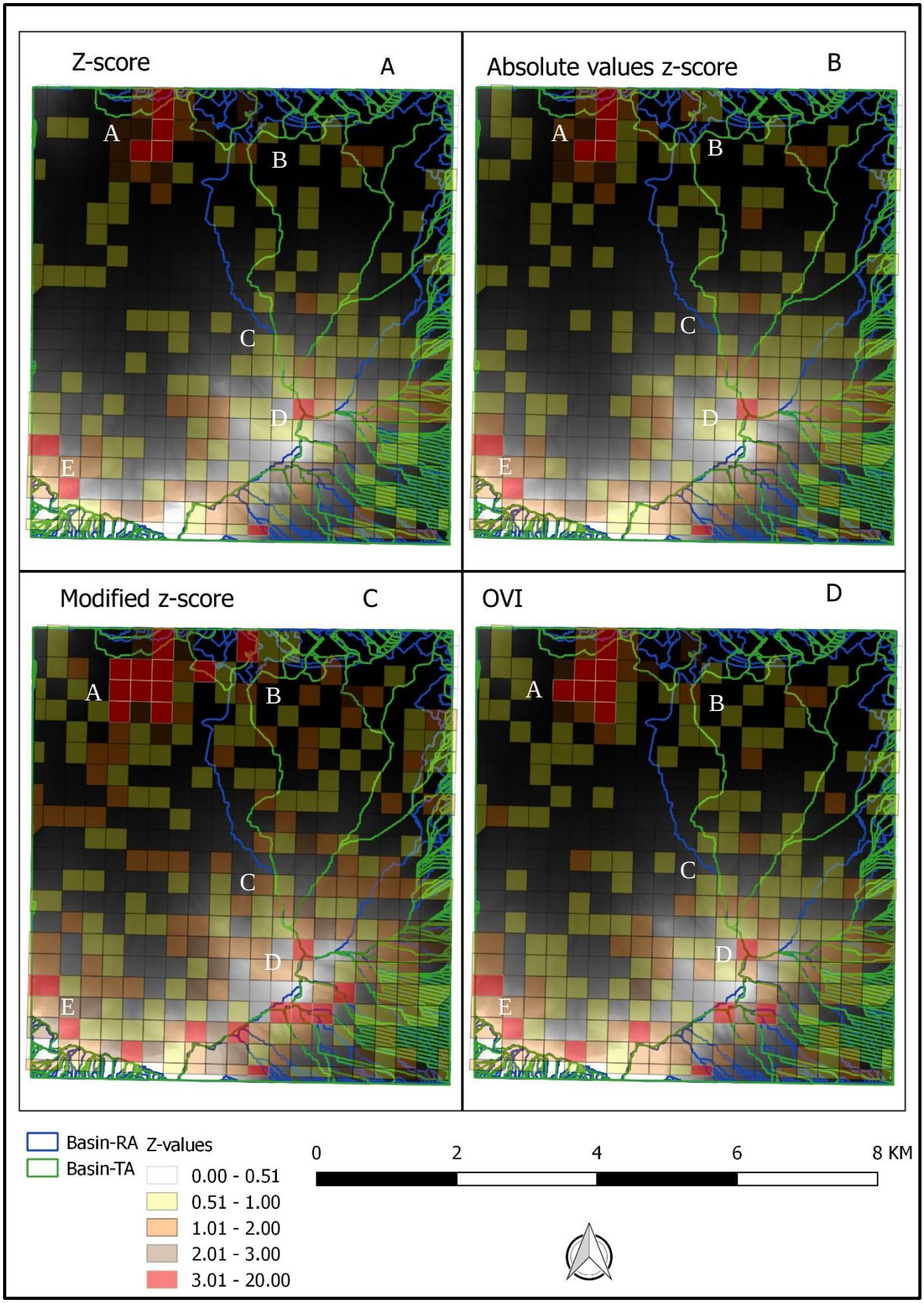


Figure 5.10 Test A z-value visualization and basin products

On first visual inspection, the visualizations for the z-score and absolute values z-score appear very similar with a larger number of outliers for the modified z-score and OVI. The histogram in Figure 5.11 confirms the observation that the z-score has the most values falling in the 0-0.5 SD z-value range, followed closely by the absolute values z-score

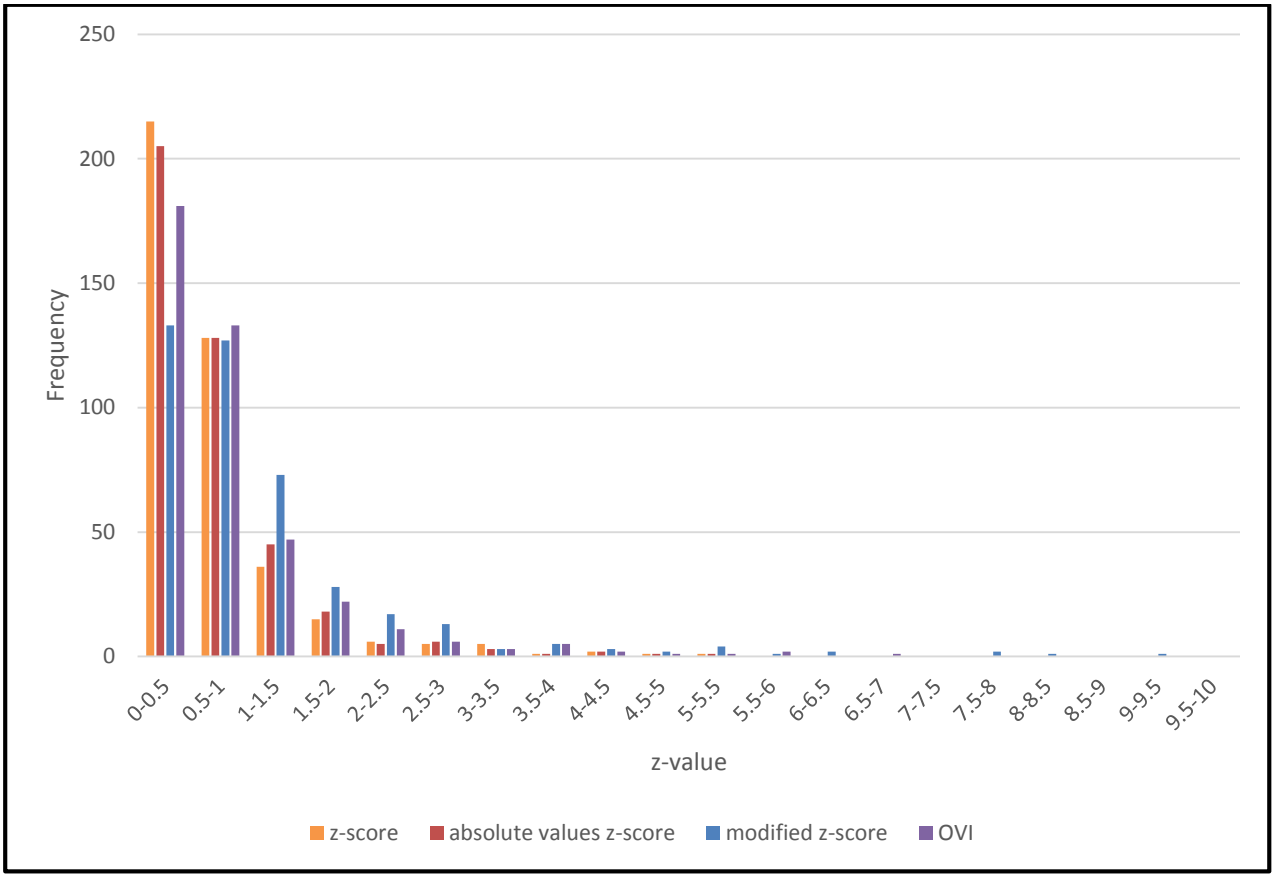


Figure 5.11 Histogram z-value based visualizations for Test A

In the 0.5-1 SD range, all metrics performed similarly, with the modified z-score flagging more values in the other ranges. The absolute values z-score and OVI, performed fairly similar in the ranges. By using the modified z-score, extreme outliers are identified in the 9-9.5 group which equates to a real difference of 20.5 m whereas the MAE for this dataset is only 2.44 m.

Figure 5.12 highlights the distributions of the four z-value based metrics in more detail. What the box plot illustrates is that the inter-quantile range for the z-score values are grouped in the 0-1 SD range. The modified z-score's interquartile range is both wider, as well as positioned slightly higher, than the other metrics. It is also clear that the z-score and absolute values z-score have very similar outliers above the upper quantile. The modified z-score has shown to be the most effective in indicating extreme outliers.

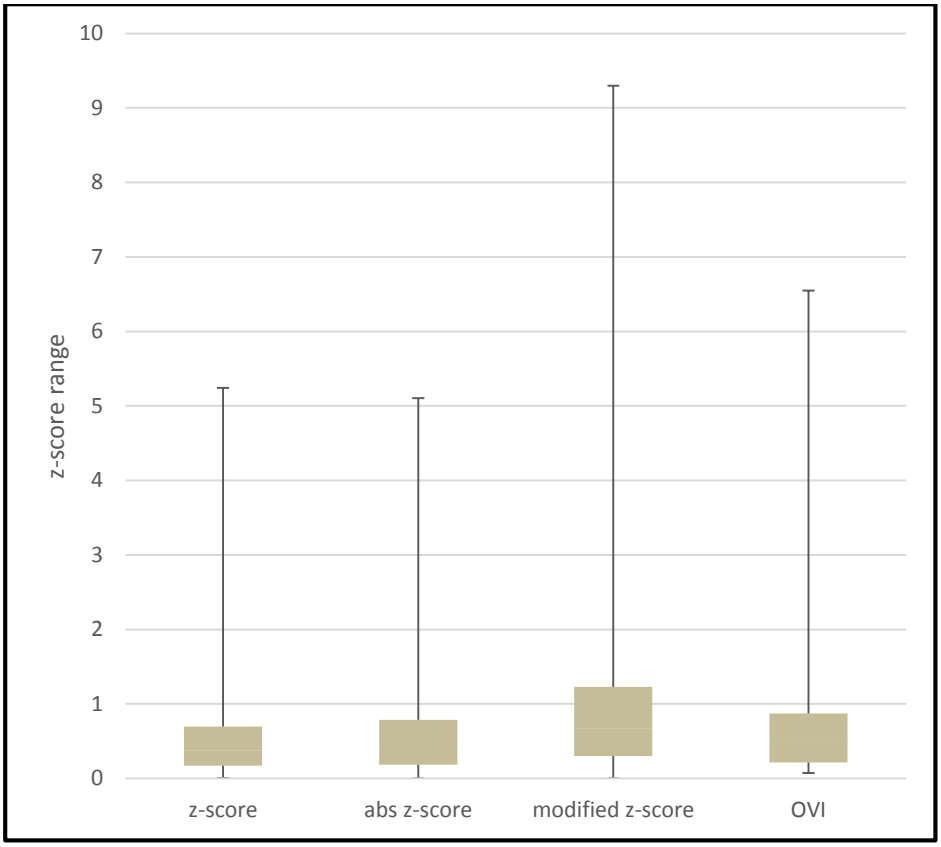


Figure 5.12 Test A box plot for z-value metrics

This is consistent with the findings from literature (e.g. Seo 2006; Iglewicz & Hoaglin 1993) that suggest, the z-score may be affected by the effect of outliers on the mean, whereas the modified z-score using the median is less affected by outliers. This can explain why the modified z-score highlights more extreme outliers.

When comparing Basin-RA and Basin-TA in Figure 5.10 it becomes clear that they do not overlap in all areas, such as areas annotated as B and C. Area C is only identified as an outlier in the modified z-score visualization. Though Area D is flagged as an extreme outlier, falling outside 95% of the data spread on all statistics, both Basin-RA and Basin-TA follow the same delineation at this point. Areas A and E are also indicated as clusters of high uncertainty on all metrics, but these areas also did not affect the basin delineations. This is in contrast with area B, where high uncertainty was indicated by the z-score and absolute values z-score and the basin delineation was affected.

5.4.2.2 Test B

Test B was then evaluated in the same manner as Test A. The first visualization that is presented here is the absolute different values using Jenks breaks in Figure 5.13.

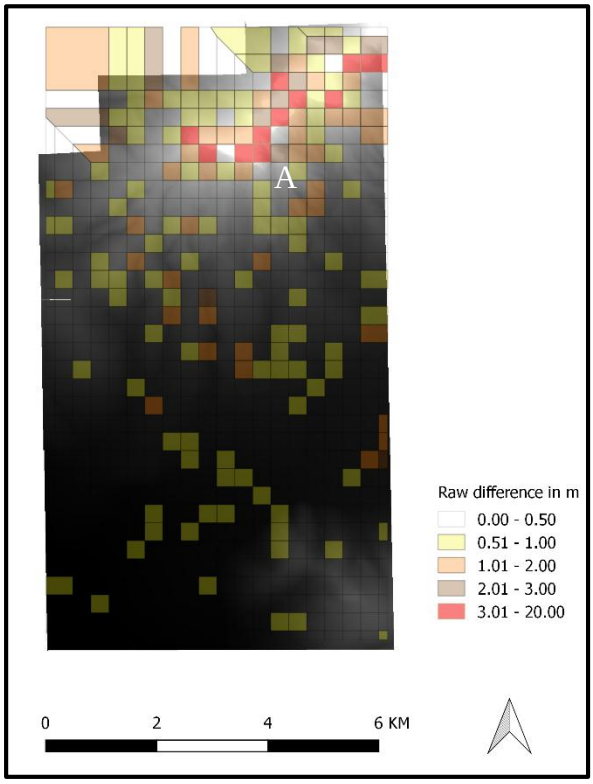


Figure 5.13 Test B absolute values uncertainty visualization

The first conclusion to be drawn from Figure 5.13, is that the maximum absolute difference between Test B and Ref B is 36.2 m. The largest differences are clustered into area A in Figure 5.13. This leads to the introduction of the statistical z-value based visualizations in Figure 5.14. All of the four metrics provided, illustrate areas of extreme uncertainty along the basin boundaries along area A, whilst the modified z-score and OVI indicate more extremes in this area. Both Basin-RB and Basin-TB follow a similar boundary at area A, whilst conversely area B is a point where the basins diverge. The z-score and modified z-score visualizations indicate area B as an area of good quality, with very moderate uncertainty in the 0.51-1.00 SD z-value range, thus still falling within 68% of the data spread.

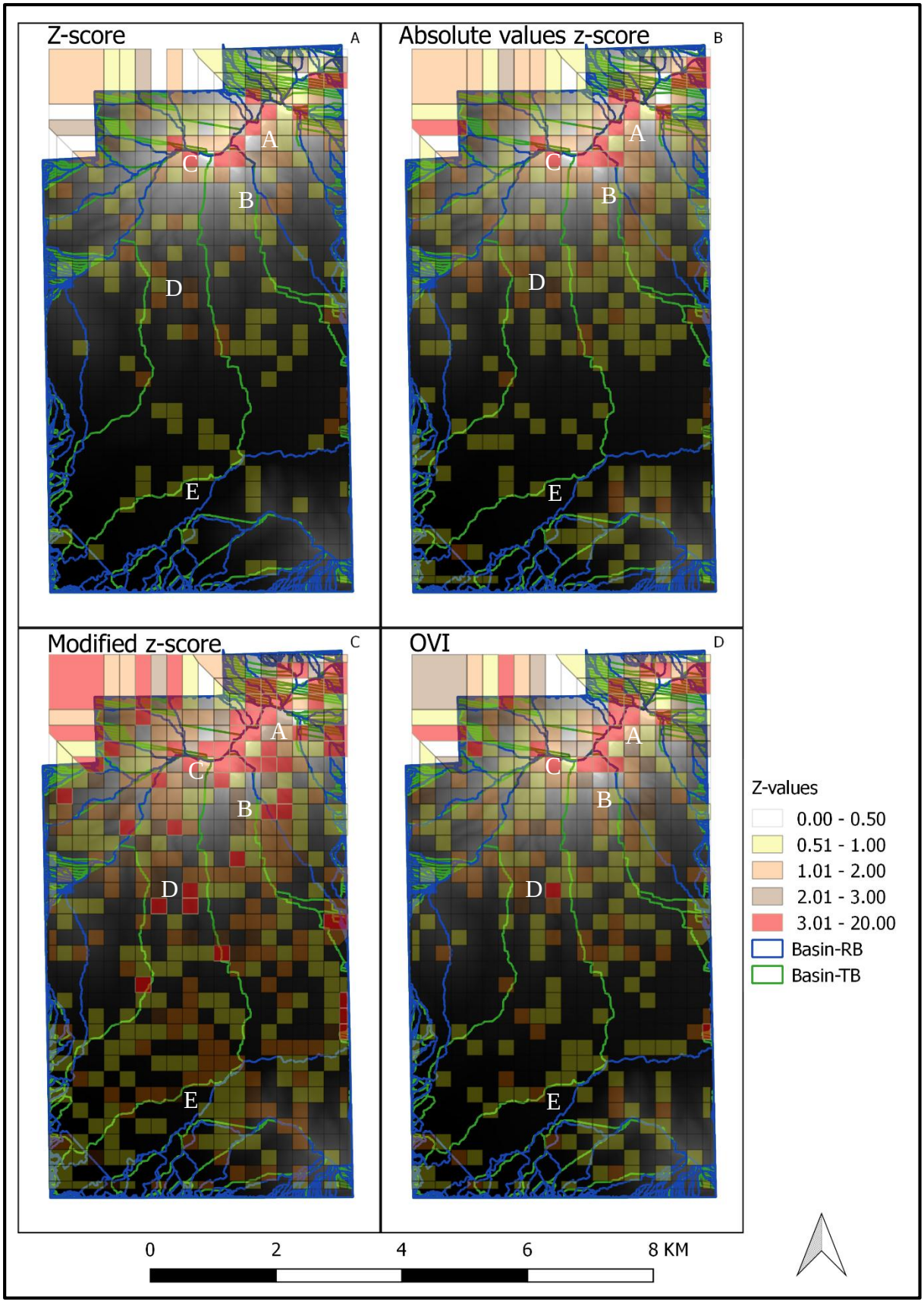


Figure 5.14 Test B z-value visualization and basin products

Basin-RB and Basin-TB also diverge at area C, on an area flagged as an extreme outlier. This divergence only occurs at area C even though both basins follow the same delineation along the highly uncertain cluster from area A down to area C. The modified z-score also flags these areas as a cluster of uncertainty, further confirming the findings from Zandbergen (2011) and Weng (2012) that uncertainty clusters in areas.

To evaluate which metric identifies the most extreme outliers, a histogram and box plot were used. Figure 5.15 shows that as with Test A, the z-score once again delineates the highest frequency of values in the 0-0.5SD z-value range (low uncertainty), with the modified z-score having the lowest frequency in the 0-0.5SD z-value range.

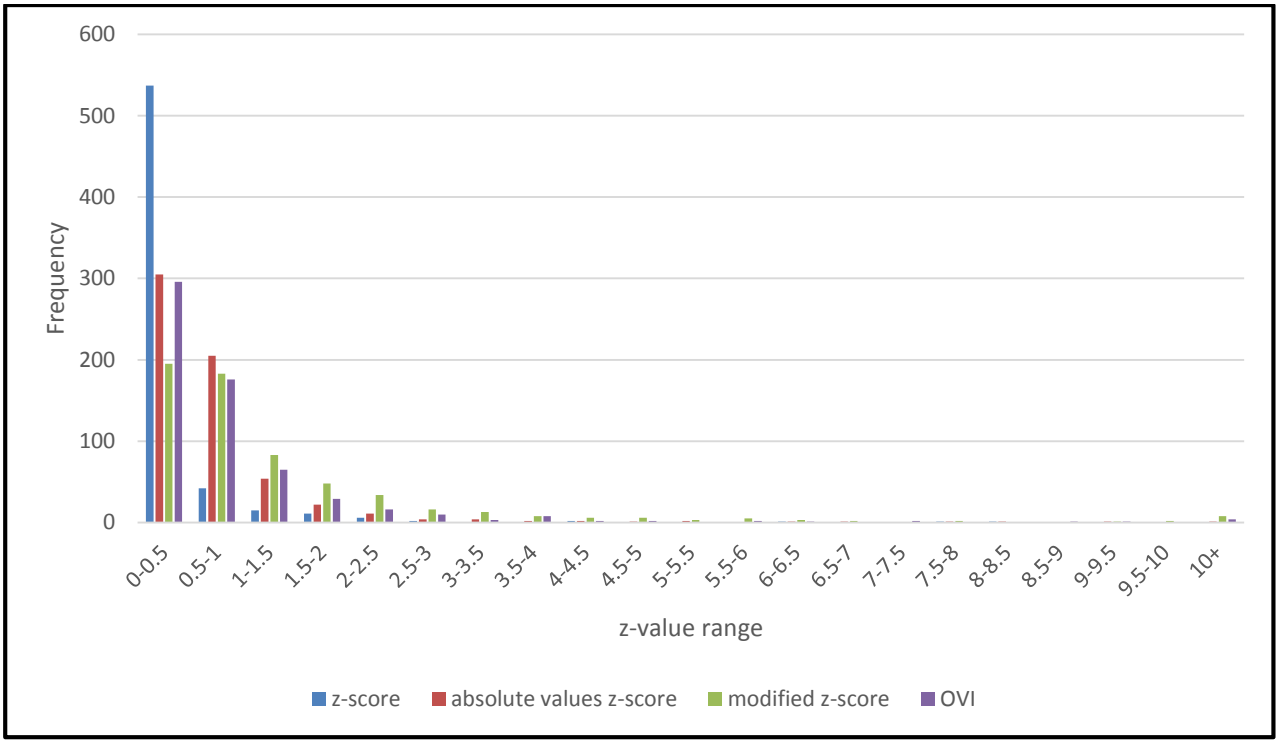


Figure 5.15 Histogram z-value based visualizations for Test B

The z-score method has the highest frequency of values within the 0.5-1 SD range, while all other metrics were grouped similarly. For this study area, as expected, approximately 70% of the values fell within one SD of the mean. The modified z-score identifies outliers better than the other uncertainty metrics. Only the modified z-score and the OVI indicated extremes past the 10+ value. Upon inspection of the attribute table, this translated to a real difference of 36.2 m. In Figure 5.16 the range of values for the four z-value based metrics are highlighted.

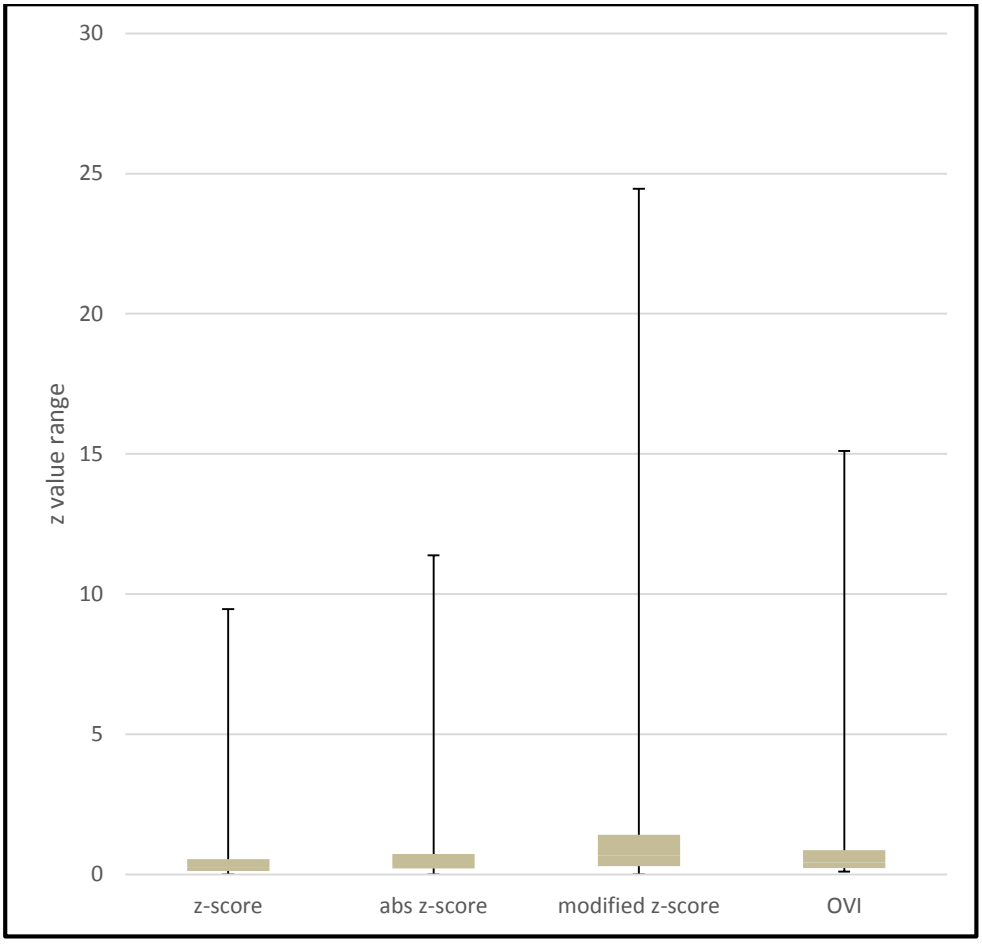


Figure 5.16 Test B box plot for z-value metrics

The absolute z-score once again shows the highest extreme, with a very high outlier compared to the other metrics. The absolute values z-score however, has a marginally bigger interquartile range than the z-score and a larger upper quantile. The OVI mirrors more of the average between the z-score and the modified z-score, whereas the modified z-score has a higher mean and interquartile range. Based on the findings from Test A in Figure 5.12 and these findings for Test B in Figure 5.16, the modified z-score is best able to identify outliers, whereas the z-score and absolute values z-score detect fewer extreme values. This confirms the influence of using the mean value in masking extremes and the ability of the modified z-score to recognize extreme outliers (Seo 2006).

5.4.3 Physical indicators of uncertainty

To investigate if there is any relationship between elevation, slope, topographic ruggedness index and higher uncertainty values, scatter plots were used. Slope represents the rate of change of elevation for each cell of a DEM (ESRI s.a.g). The topographic ruggedness index is a measure that quantifies the total altitude change across a given area (Liang, Kang &

Pettorelli 2016). The z-score was chosen as the only metric to report for these evaluations, as the modified z-score performed similarly with at most a 0.02 difference in R^2 between the two metrics.

5.4.3.1 Elevation

The first measure evaluated, was the relationship between z-score and elevation. In this section both Test A and Test B will be discussed together, with conclusions drawn from them for each element. The plot for Test A seen in Figure 5.17, shows that the trend line between z-score and elevation indicates a weak or no relationship between uncertainty (z-score) and elevation.

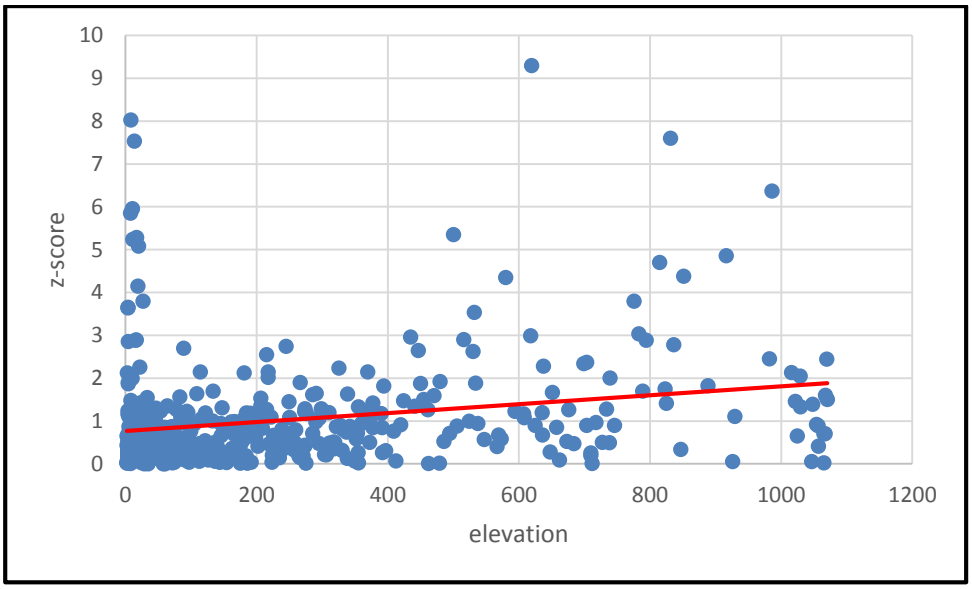


Figure 5.17 Test A z-score / elevation relationship

What needs to be noted, is that at low elevation there are some very high z-values, indicating high uncertainty. The Uview visualization with the hillshade as backdrop (Figure 5.18) for Test A, relates the cluster of high uncertainty at low elevation at area A, with the high values seen in Figure 5.17.

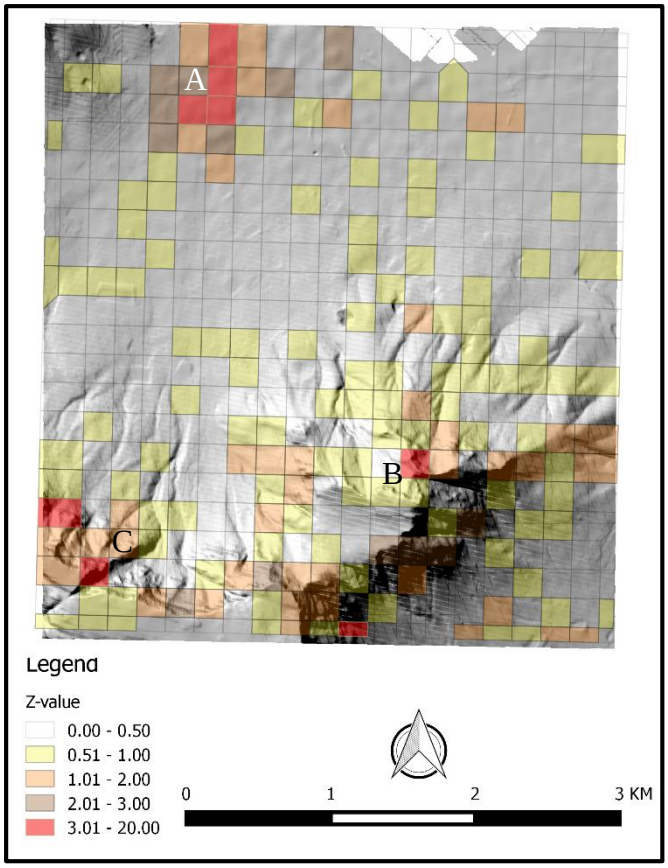


Figure 5.18 Test A hillshade with z-score uncertainty overlay

For Test A, there is not a clear relationship between elevation and uncertainty. Further investigation of the plot in Figure 5.17 also shows extreme z-values (z>3) occurring at elevations greater than 600 m, as well as at elevation around 20 m.

The plot for Test B can be seen in Figure 5.19, showing that all outliers with a z-score of above 10 occur at an elevation of above 650 m, associated with the cluster of uncertainty values indicated in the Uview visualization in Figure 5.20 at area A.

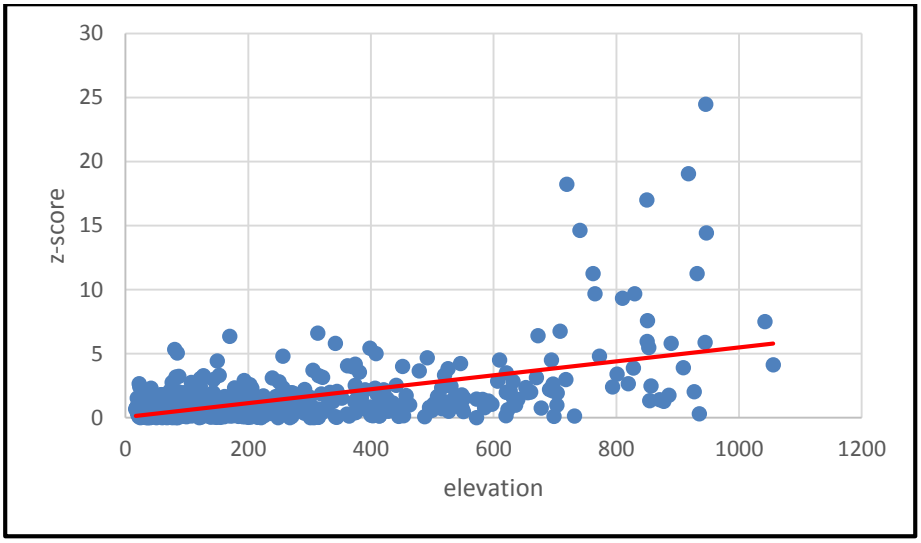


Figure 5.19 Test B z-score / elevation relationship

When looking at Figure 5.20 there is only one cluster of uncertainty that runs at area A. Unlike with Test A (high and low lying clusters of high uncertainty) there is only one cluster of high uncertainty here, running along the ridge at area A.

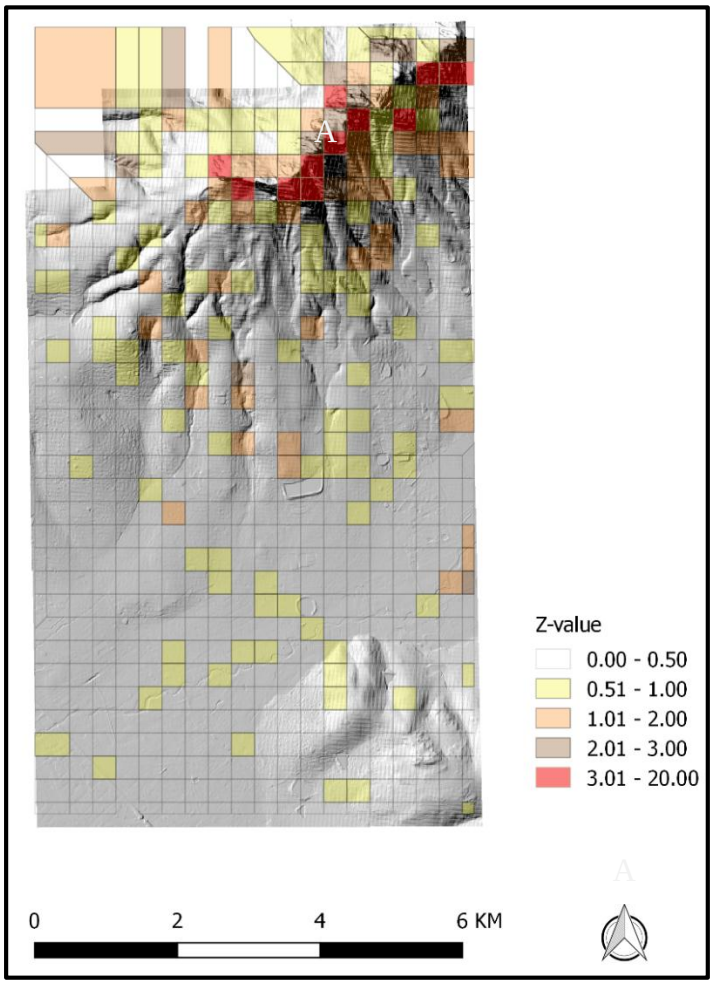


Figure 5.20 Test B hillshade with z-score uncertainty overlay

From both Figure 5.19 and Figure 5.20, it is clear that the extreme outliers in this area are present mostly at areas of higher elevation within the dataset. Elevation in Test B thus has a stronger relationship to uncertainty than Test A. As it is possible for uncertainty and extremes to occur at any elevation as seen in Test A in Figure 5.17, it is thus not a definitive factor to consider for where to expect higher uncertainty.

5.4.3.2 Slope

Slope was also investigated for correlation with uncertainty. As with elevation for Test A, there are large z-values present even at low slope values (Figure 5.21).

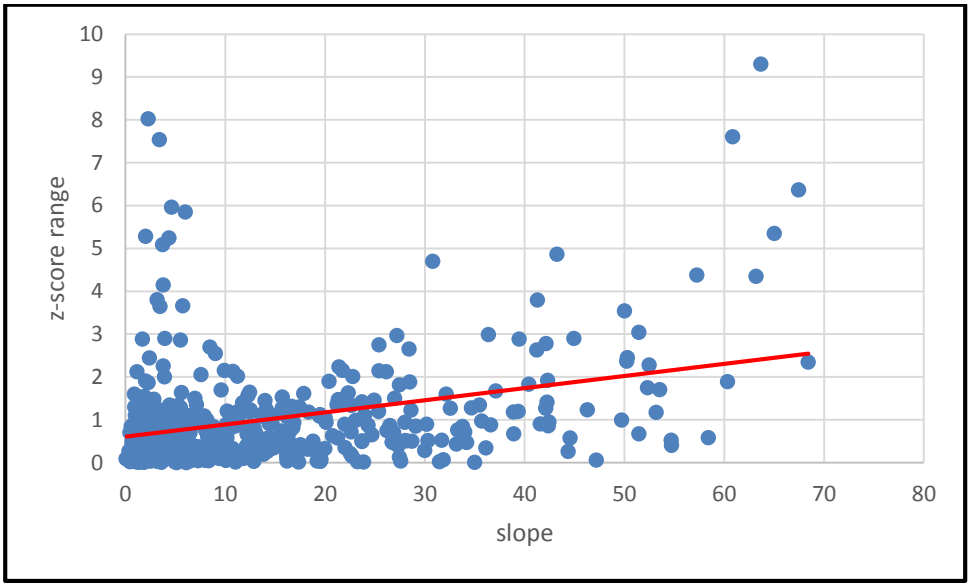


Figure 5.21 Test A z-score / slope relationship

High z-scores are found both at higher, steep slopes, as well as at the relatively low flat area that has a cluster of uncertainty at area A in Figure 5.18. Figure 5.21 also suggests that above a slope of 60, the data quality drops off, as only one of the seven measured points have a z-score below 2. From the elevation results, it is thus expected that a steeper slope in Test B will relate better to higher uncertainty than in Test A. Figure 5.22 confirms this: below a slope of 20 there are only five values with a z-score value of above five, however above a slope of 40, more than half of the evaluated points have a z-score of more than 5 indicating extreme outliers.

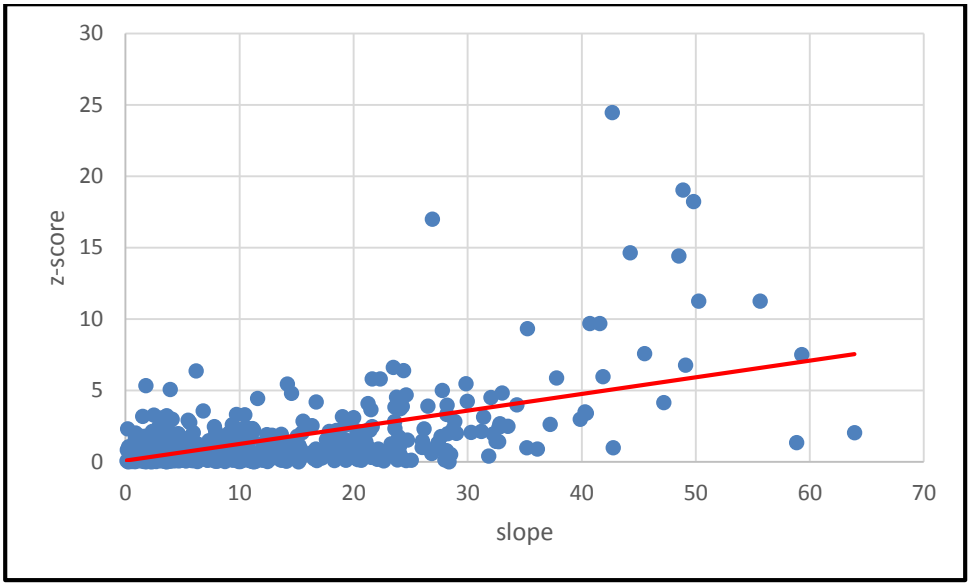


Figure 5.22 Test B z-score / slope relationship

Slope has a better correlation (0.32 R squared as opposed 0.29 R squared for elevation for Test B) with uncertainty than elevation. It also indicates that at the resolution of this analysis (5 m), there is a slope steepness above which data quality starts to drop off. One more index was used to test for correlation with uncertainty.

5.4.3.3 Topographic ruggedness index

The topographic ruggedness index was used to determine if an uneven surface has any relation to uncertainty (higher z-score values). Figure 5.23 shows a similar result as with the elevation and slope, that there is a lot of uncertainty at both low and high values in Test A.

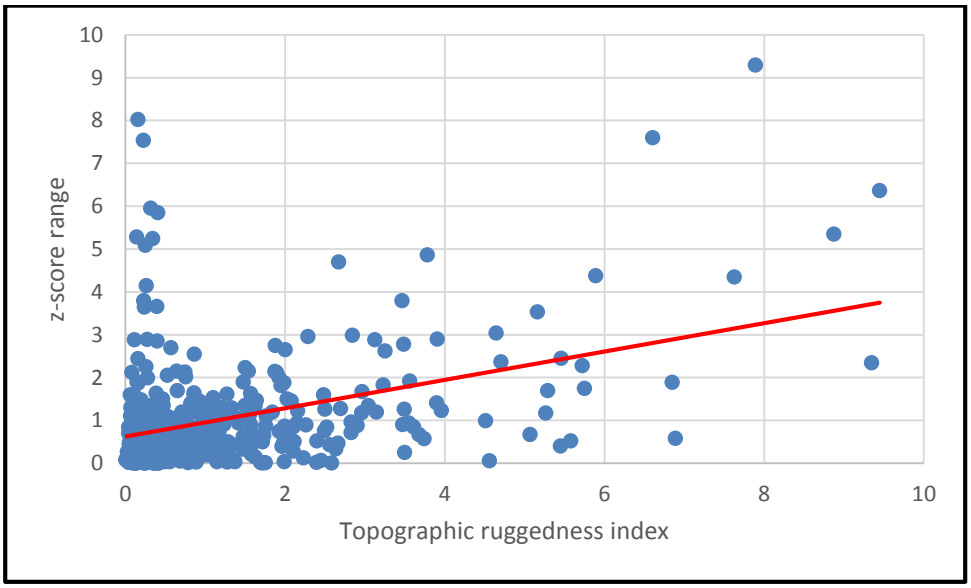


Figure 5.23 Test A z-score / ruggedness index relationship

With the topographic ruggedness index however, there is a better relationship between uneven surfaces and uncertainty. The extreme outliers seen in Figure 5.23 at low ruggedness, are still present as with the other two measures. Above a ruggedness index of six, data quality starts to decrease, as only two out of eight points (75% probability of a z-score above 2) are located below a z-score of two with one being very close to two. Test B indicates similar results in Figure 5.24.

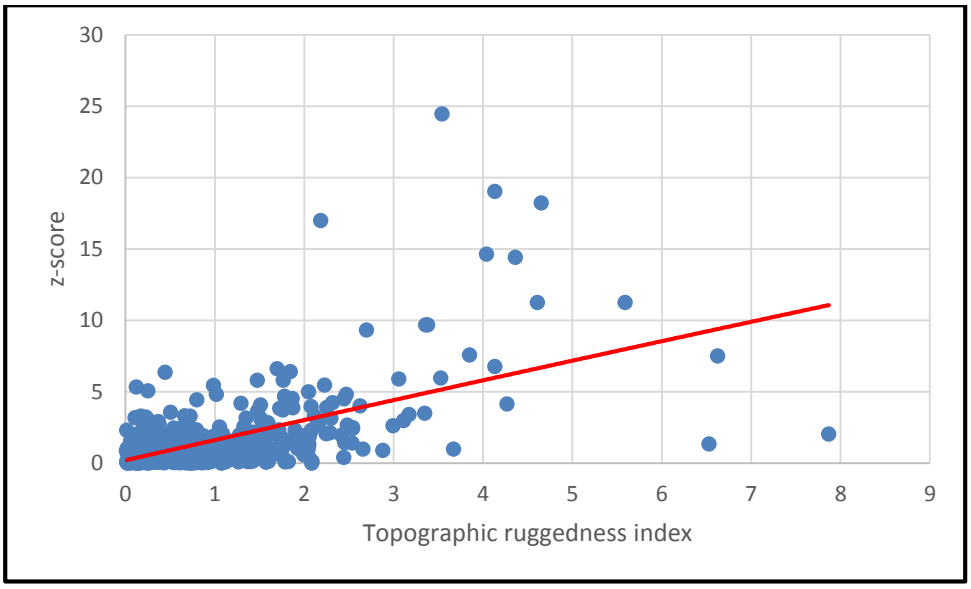


Figure 5.24 Test B z-score / ruggedness index relationship

Points in Test B above a ruggedness index of four (Figure 5.24), all reflected z-score values above two with the exception of two points out of 11, equating to an 81% probability that an area above a ruggedness index of two will have a z-score of two and above and thus be an extreme outlier. At a ruggedness index of four for Test A, there is a 54% probability that a z-score above two may be found. Thus a terrain ruggedness index of four or higher is an indicator of potential higher uncertainty in the DEM when compared with less rugged areas.

For DEMs of this resolution (5 m), ruggedness can therefore be linked to the possibility of uncertainty, which is in agreement with the findings of Weng (2012). Uview with its four statistical visualizations, highlights uncertainty in all areas both rugged and smooth. Though the z-score and absolute values z-score performed similarly when compared in box plots, on a histogram the z-score indicated fewer extreme values. All statistical metrics produce usable results based on similar performance of the two outliers. Deviations for watershed models between Basin-RA – Basin-TA and Basin-RB – Basin-TB appeared to show a weak correlation with areas of higher uncertainty identified by Uview. This confirms the user's

responsibility in considering the quality of the data that is used and the possible sensitivity that individual models may demonstrate to data.

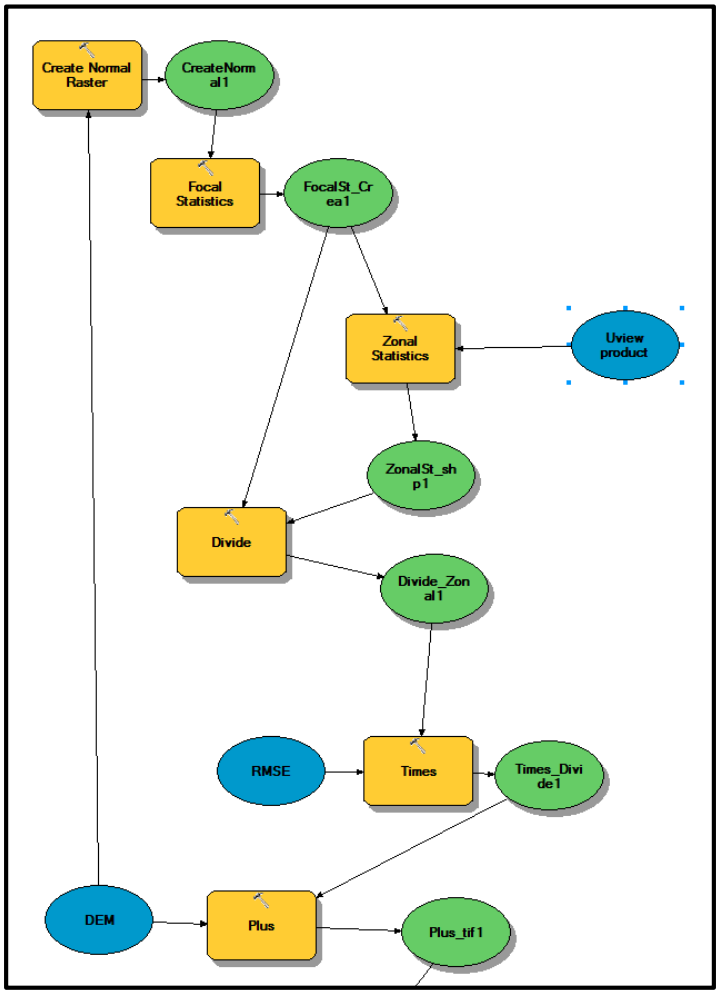
5.4.4 Which visualization?

Uview provides both statistics and a visualization for spatial communication of uncertainty in geospatial data. Uview makes use of the z-score, a statistic that standardizes the data based on the standard deviation (SD) from the mean of the data. The mean of a dataset is however sensitive to outliers. Therefore, the modified z-score, which tries to remove the effect of outliers by using the median absolute deviation (MAD) and the median instead of the SD and the mean respectively was also chosen as a metric. Although the absolute difference is the easiest to understand, it is not normalized and thus does not give a statistical difference, or the ability to compare between visualizations. Visualizing the distribution of uncertainty demonstrates that uncertainty in DEMs occur clustered, but at random locations as well (Zandbergen 2011; Weng 2012). The quality of the DEM has a great effect on the derived product (Zhao et al. 2009) as seen by the difference between the test and reference basin products. It was also found, that large deviations may not have an effect on the derived product, but small areas of uncertainty at key locations may have an effect on the derived product. Uview visualizations combined with traditional statistics such as RMSE and MAE, communicates the quality of the data to users and producers in the most powerful way, namely in the visual way (Bostrom, Anselin & Farris 2008).

5.5 DEM MODELLING

The second evaluation demonstrates the effect of uncertainty in model input data on model output, by introducing random error into the input data through Monte Carlo simulation (Zandbergen 2011). Uncertainty is introduced into the 5 m SUDEM by selecting the three categories of highest uncertainty identified by Uview and using them as focus areas for randomization (Zandbergen 2011). A watershed is then calculated based on the newly modelled dataset and confidence intervals for the boundaries are developed. The model is based on random error occurring anywhere within the DEM, but also on special clusters where uncertainty is more likely to occur. Zandbergen (2011) and Fisher (1992) agree, that error in DEMs are both spatially correlated and random in nature. The model is built in ArcMap and is illustrated in Figure 5.25. In the model, a raster is created with random values assigned based on a normal distribution. Cells populated are negative and positive with a resulting mean of zero and a standard deviation of one. The Focal Statistics tool is used to

introduce spatial autocorrelation and cluster errors into areas where the random error is averaged. As focal statistics will change the standard deviation, zonal statistics is used to bring the standard deviation of the new raster back to one. The Uview uncertainty layer is used as input zonal statistics layer for this calculation. The error raster is then multiplied by the RMSE (that is supplied by the user based on the accuracy of the original DEM), to create a spatially autocorrelated error DEM with a mean of zero and the input RMSE as standard deviation (Zandbergen 2011).



Adapted: Zandbergen 2011

Figure 5.25 Error simulation

The Error DEM (created through the method explained in Figure 5.25) is then added to the input SUDEM, which is used to delineate a watershed, following the process depicted in Figure 5.8. This process is performed multiple times and the resultant delineations are superimposed over each other to create a probability map based on how many times a line is defined as a catchment boundary as a percentage of the total number of runs. This watershed

probability dataset is then compared to the Uview z-score visualization, to identify any correlation to areas of higher uncertainty. This section of the case study demonstrates what can occur if an accuracy assessment is not consulted and data quality is assumed to be high. This assumption of good data quality was the case for 50% of those respondents to the survey with more than 10 years' experience and 70% of those with less experience (Chapter 3).

5.6 ERRORS IN WATERSHED MODELS

Once DEM modelling was performed (Figure 5.25), the same procedure was followed as suggested by TUFTS University (2012) for delineating watershed models. The error model was run 100 times to create 100 delineations. The polygon shapefiles were then converted to line files and assigned a value of one. These were then converted to raster datasets and combined using ArcMap. The final output from the model is a raster where the cell values represent the number of times the watershed boundary has been modelled. Thus, at cells where the boundaries occur 100 out of a potential 100 times, the probability of the watershed boundary occurring here is 100%. The model result provides an indication of the uncertainty in boundary delineation. Test A was used for this model, as it showed both high uncertainty at low and high elevations. The analysis compared the occurrence of low probability watershed delineation on the map with areas where Uview highlighted the most uncertainty between the Test A and Ref A. Test A was also modified to match Ref A in all areas of falling in high uncertainty, as identified in the Uview z-score product (areas in z-score of 1.01-). This edited version of the Test A DEM will from now on be referred to as Test A-Cor and described as partially corrected. Probability vs. Uview visualization will be discussed as occurs in Figure 5.26.

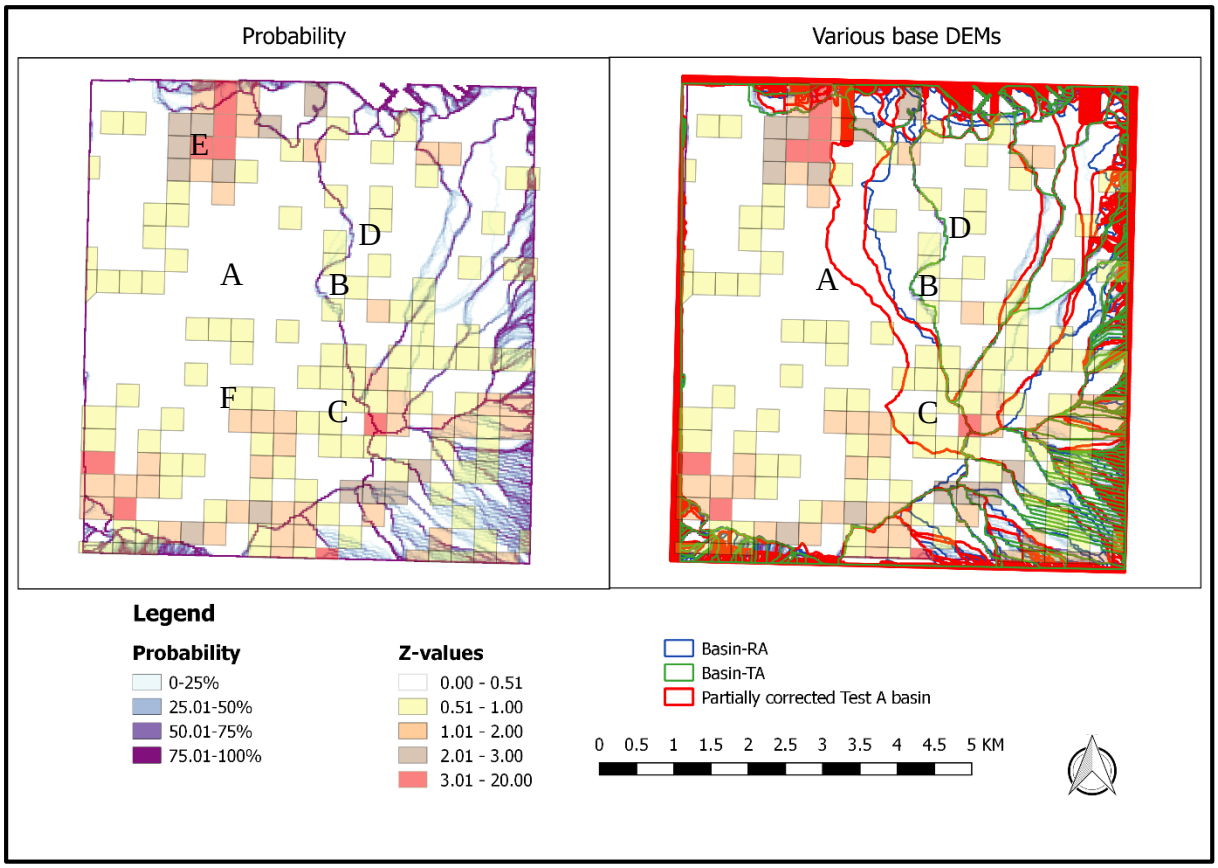


Figure 5.26 Probability and partially corrected Test A DEM

The way the probability map is setup, is such that areas with a higher probability are in a darker shade of blue and the areas with lower probability are in a lighter shade of blue.

As can be seen in Figure 5.26 the probability map areas of higher uncertainty do not always correspond with areas having lower probability. There is one large watershed / basin delineated that with E, A, B and C in with high Uview z-score values of 2.01+ clustered around E does not affect the outcomes of the basin probability. At area F, another cluster of uncertainty occurs that also has no directly observable impact on the probability of the basin. These clusters however confirm the findings of Zandbergen (2011) and Weng (2012) that uncertainty happens in clusters. At location D however, an area with some uncertainty albeit only in the 0.51-1.00 z-score category, there is a lower probability at these areas of uncertainty, however this indicates further that at key areas even a small deviation can lead to a different derived product.

When looking at the second image in Figure 5.26, three derived basins can be seen. Basin-RA, delineated from Ref A and partly hidden behind the other basins, will be treated as the reference data for use as ground truth. Basin-TA is the basin as modelled from Test A, and

the partially corrected Test A basin is named Basin A-Cor. When one compares Basin-TA and Basin-RA, they do not diverge from each other at areas identified as high uncertainty by the high z-score in Test A, but instead they diverge downstream or at areas of relative low z-score. When comparing Basin A-Cor with both Basin-TA and Basin-RA, one finds a rather different product, especially at A in the various base DEMs map in Figure 5.26. Basin A-Cor closely resembles Basin-RA at the diversion from Basin-TA at B, Basin A-Cor however deviates from both the other basins at area C, creating an extra basin. The split between Basin A-Cor and Basin-TA at D is interesting, because if one compares it with the probability map at D, it follows one of the lower probability basins. When one compares all the three basins, one comes to the conclusion, that the accuracy of a DEM has a large effect on the output. This relates to the work of Zhao et al. (2009) who found that elevation and corresponding difference from a reference was the only thing that affected DEM derived products more than resolution. As can be seen with the difference between the Basin-TA and Basin-RA even when one employs a Monte Carlo simulation to derive a probability test and introduce random errors to the magnitude of its measured error, Basin-TA may still not create a delineation similar to that of the Basin-RA that was treated as more accurate. Partially correcting Test A (Test A-Cor) also produced a different basin. This again confirms the argument that spatial data users have to understand the quality of the data they are working with, as it has a direct influence on the products generated.

The overall conclusion reached is that due to the nature of basin models and DEM derived products, the quality of the input DEM will have a direct effect on the quality of the product. Therefore, whenever a DEM is used as an input device, care has to be taken not only of the resolution of the DEM, but also the quality of the DEM (Weng 2012; Zandbergen 2011; Zhao et al. 2009). Watershed/basin models appear to be sensitive to data quality in the broader array, and even minor to small deviations may produce different output products as seen in Figure 5.26. Even when the large discrepancies were corrected, the resultant basin did not follow the course of the more accurately rated LiDAR dataset. Simulation, although useful, may also still not provide the full picture, but it is one method of understanding the potential weakness of the input data used.

Uview thus provides a tool where users and producers can explore the quality of their product, both statistically through the accuracy assessment statistics that are provided, but also visually through the visualizations provided. It is also imperative that those using

datasets be cognisant of how sensitive the processing they aim to do with the data is to data quality.

5.7 CHAPTER FINDINGS

Uview is a QGIS plugin which makes it one of the easiest and most accessible tools for GIS professionals, especially those who are already using QGIS. It can be easily incorporated into a standardized workflow, and requires a minimum of two inputs and a maximum of three. Uview provides the statistics used in traditional accuracy assessments of continuous raster data, together with a visualization to communicate the element that goes missing in most accuracy assessments. This is the spatial element of data quality, as the spatial nature is what distinguishes geospatial data from any other form of data.

When using the Uview product to evaluate the products of two DEM derived watershed models, one treated as a high-quality reference and the other as a dataset of lesser quality, Uview provided a visualization as to how the uncertainty was spatially distributed and how the original DEM visualized uncertainty correlated to the difference in the watershed products. When comparing the deviations of the derived basins from the two datasets, it was clear that nonconformities did not always occur at areas of high difference between the input DEMs. Deviation may occur at areas of low deviation and thus it is important for those working with these products and their derived products to understand how the watershed model works. It thus highlights that knowing your data is not only important, but it is also necessary to know the techniques used and the purpose of the data. It was further found, that for DEMs there is a correlation between terrain ruggedness and uncertainty. The more rugged terrain becomes, the higher the chance of uncertainty, especially above a topographical ruggedness index of four. Uncertainty is however not limited to any ruggedness index as demonstrated in Figure 5.23. Elevation and slope were also tested for the correlation with uncertainty, but the topographic ruggedness index provided the best correlation. With the simulation that was run, it was also shown that even simulations may not fully show areas of weakness in data.

Uview is limited to continuous data at present and still requires reference data. A lot of work still needs to be done on uncertainty visualization tools, especially those that work on the raw data and not on the products thereof that can have Monte Carlo simulations, or similar simulations produced for quality visualization. Monte Carlo simulation has also shown that,

depending on how sensitive the model is to data quality, it may still provide a different result to what is achieved with better quality data even when resolution is kept at a constant.

Uview was thus found to be a tool that can be used not only for accuracy assessments and visualization, but also for further researching the sensitivities of models to different types of data discrepancies.

CHAPTER 6 QUALITATIVE EVALUATION OF UVIEW

Chapter 6 is a qualitative evaluation of Uview; it addresses the last sub-section of Task 3; and relates to the fifth objective (see Figure 6.1). For this chapter, some of the previous data has been used, namely the Uview product relating to Test A and Ref A in Chapter 5.

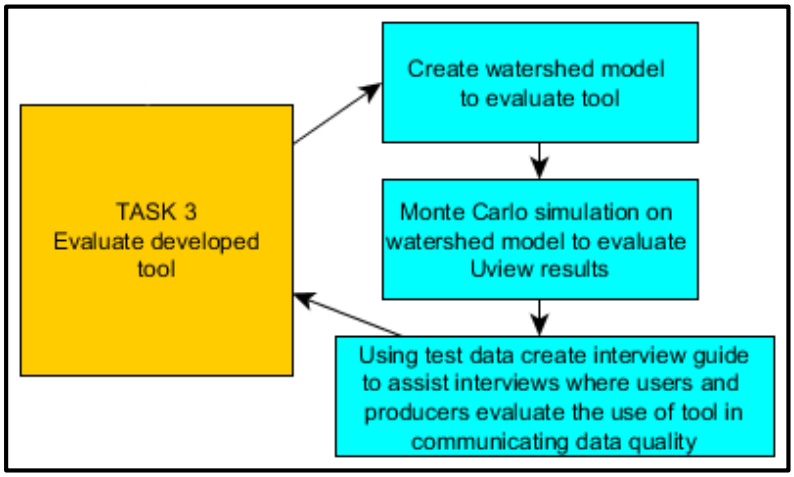


Figure 6.1 Task 3

A qualitative evaluation of Uview was conducted through interviews with potential users of Uview. Two types of qualitative evaluation exist, namely deductive analysis and inductive analysis (Elo & Kyngäs 2008). Deductive analysis starts with a hypothesis or theory which is already held or known, and data is then collected to test this hypothesis or theory. During inductive analysis, on the other hand, a researcher will start by collecting a mass of data about their field of interest and, once completed, the data will be evaluated and theories will be developed that attempt to explain the trends found within the data (Asaka 2016; Blackstone 2016; Elo & Kyngäs 2008). For this evaluation of Uview, the deductive approach was chosen as the tool was already available, as well as some existing theories around the use of uncertainty visualization. Three theories were investigated using the interview process:

- The tool produces a visualization that adds value to the understanding of statistical methods of evaluation (Perer & Shneiderman 2009);
- Producers think a visualization may reduce the perceived quality of the dataset, based on the findings of Kinkeldey & Schiewe (2014);
- There is no set method by which producers and users deal with uncertainty in datasets, related to the finding 44s of Tegtmeier et al. (2007).

Ethical clearance was obtained for the interviews from the Stellenbosch Research Ethics Committee (REC) see Appendix C for informed consent as set out and approved by the REC (for this section of the research). The full interview guide, containing the questions can be found in Appendix D.

In total, twelve interviews were conducted with four current academic staff, four working professionals and four recently graduated students from various universities.

6.1 EVALUATORS AND RESPONSES

Eligible evaluators were selected for interview based on their involvement and skill with geospatial data, using the same channels as described in Chapter 3. The Geo-Information Society of South Africa (GISSA) assisted by posting an advert on their Western Cape Facebook page for those interested in assisting as evaluators of Uview. Preliminary research findings were presented at a quarterly GISSA Western Cape 2016 meeting (titled: Uncertainty and visualization), with a request for evaluators. In addition, the Open Source Geospatial Foundation's (OSGeo) Africa chapter issued an email invitation for people willing to aid in the evaluation of the research. Lastly, three lecturers at the researcher's university were requested to participate as evaluators. Although the response to these calls was low, the researcher finally proceeded with a total of 12 evaluators.

In-depth interviews were scheduled with these 12 evaluators, in line with similar work by Kinkeldey & Schiewe (2014), who also only used the same number for their qualitative study on the management of uncertainty. In the professional environment of geospatial study it is not easy to find evaluators, especially when it involves taking them away from their work time. However, Baruch & Holtom (2008) and Rogelberg & Stanton (2007) mentioned that in areas of little knowledge, even a small response rate can deliver significant results and open space for further and new research, once some new knowledge has been created. All twelve evaluators selected have either recently completed studies in geography or geoscience, or are actively working in the geospatial industry and education. These evaluators were placed in three groups: *Graduates*, *Industry* and *Academics* (hereafter italicized for clarity).

Interviews were arranged at times and places convenient to the evaluator. The researcher conducted all interviews so as to provide the opportunity to gain maximum information from the respondent evaluators. All questions were open-ended, providing the best scope to take full advantage of any new direction that may be given and obtain the most relevant

information from the evaluator. Most interviews were conducted face to face, but a few were conducted via Skype, as the interviewer could not travel to the location of the evaluator.

6.1.1 Theories, themes and questions

Three theory linked themes were developed into six semi-structured interview questions, with a seventh question on the software tool usability (see Table 6.1):

Table 6.1 Themes, theories and questions for interviewed evaluators

Theory linked themes	Related questions
Effectiveness: - Understanding - Statistics vs. visualization.	(1) What do you understand by uncertainty in geospatial data? (3) How do you feel about the data after a visualization?
Perception: - Visualization as an aid in understanding uncertainty. - Visualization detracts from perceived data quality?	(3) How do you feel about the data after a visualization? (6) Do you feel that the visualization of uncertainty would degrade the perceived value of a dataset?
Management: - How uncertainty is managed.	(2) How is uncertainty managed? (4) Do you consider visualization of uncertainty would aid in management? (5) How do you consider uncertainty should be communicated to end users and would visualization aid in this?
The software tool	(7) Do you have any suggestions or remarks about the usability of the developed tool?

Interviews were designed to firstly elicit from each evaluator their understanding of uncertainty and the management of uncertainty in their workflow. Then, global statistics for a

dataset (normally supplied as metadata) were presented, after which Uview was demonstrated on this dataset for comparison with the statistics. This led to examining whether the use of visualization in Uview created a different understanding of the data quality, and whether it would improve the current management of uncertainty. The question was then raised as to how uncertainty is currently communicated by producers to the end users, together with the potential role of visualization in future communication. The aim here was to determine whether visualization would aid positively in the communication of uncertainty downstream, or whether it would potentially have a negative impact on the perceived data quality. Evaluators were also invited to comment on the usability of Uview to identify potential shortcomings, as well as to provide suggestions for its improvement.

For analysis purposes, interview responses were grouped into the respective themes developed (see Table 6.1 above) to test the three theories.

6.1.2 Responses

Responses from the *Graduates*, *Industry* and *Academics* groups were measured against, adding value to understanding statistical measures of uncertainty in spatial data, perceptions of data quality and issues of workflow management in communicating uncertainty. In addition, suggestions for improvement to this visualization tool are presented.

6.1.2.1 Effectiveness of statistical vs. visualized uncertainty

Question one established the evaluator's understanding of uncertainty, thereby providing a baseline of how well they understood statistical methods, and served to measure how closely they related uncertainty with statistics provided in metadata. To determine whether visual methods of uncertainty representation were easier to understand than statistics, a dataset and its quality assessment statistics were first introduced. Visualization of uncertainty measures for the dataset were then presented in Uview, to see if this led to a better understanding of the statistics and the quality of the dataset (Question 3: Table 6.1). As demonstrated by the survey reported in Chapter 3, all evaluators in this survey were aware of uncertainty. The *Academics* group provided the best textbook definitions for uncertainty, in agreement with Kinkeldey & Schiewe (2014), by defining uncertainty as a fuzzy concept dependent on the particular task. The definition of uncertainty by the *Graduates* and *Industry* groups ranged from "the difference between a dataset and what it presents" to "the unknown inaccuracies that we find within datasets." Though more experienced professionals (from *Academic* and

Industry groups) agreed that both data quality and uncertainty depended on the job at hand, no-one directly mentioned statistics as a key factor when thinking about uncertainty.

After the dataset's statistics were introduced, *Academics* and *Industry* groups (with one exception) agreed that although the dataset was of good quality, the dataset use (purpose) determined if it was 'good enough'. Though the quality may be good enough for a line of sight project, a sensitive flood model may require a better quality dataset. The *Industry* group exception indicated that the required quality would depend on the client requirements, which dataset the client supplied or which was easily sourced nationally. If data is supplied by the client, dataset quality is not generally considered in great detail, but is simply incorporated into the workflow. However, these responses did not indicate that the impacts of poorer data quality on resulting products from geoprocessing or modelling, are communicated to the client. *Graduates* all considered the dataset to be of good quality based on the statistics. One *Graduate* group member indicated that not much focus was placed on data quality in their current work environment, and that the geospatial products generated were often supplied without accuracy assessment information.

When Uview was demonstrated to the evaluators, they all agreed that visualization provides insight into spatial location of uncertainty and, together with the statistics, provides a better understanding of the dataset quality. Concerns voiced by some *Industry* and *Academics* group members related to the definition of uncertainty, specifically on the error *vs.* uncertainty notion and the uncertainty metrics used. Uview represents uncertainty as the difference between a dataset and the phenomena that it represents (Longley et al. 2005), extrapolated to the extent of a Voronoi polygon created around the point where uncertainty is measured (based on Tobler's first law of geography that near things are more related than more distant things (also known as spatial autocorrelation (Sui 2003). This also relates to the definition that uncertainty is when error is not known as the areas are not known, but inferences are made about them due to proximity to a known area (MacEachren et al. 2005). Quality of the data over the whole polygon is represented from inferences based on the data at the closest measured point. Therefore, Uview combines both definitions (Longley et al. 2005; MacEachren et al. 2005) in its visualization. As the statistics generally provided in an accuracy assessment are global and not linked to spatiality, uncertainty linked to location is not reported. There was thus consensus, that combining statistics with visualization would communicate quality in geospatial data more effectively, especially with local uncertainty throughout the dataset being seen.

From these findings and the findings in Chapter 3, it can be concluded that most geospatial data users are aware of uncertainty in data. However, the definition of uncertainty remains a fuzzy notion, with no clear consensus between different respondent groups. Furthermore, statistics are a global representation of uncertainty, which is a factor not always considered. Not everyone looks at the statistics or fully understands what they imply, but the downside is that statistics can also be limiting, especially when mean statistics are given, which are prone to both the effects of outliers and averaging effects over a large area (Seo 2006). Thus, visualization combined with global statistics, can aid in bringing about a better understanding of data quality, both for producers of data as well as for users of data products.

6.1.2.2 Perception of visualized uncertainty on end users

Results of this inquiry relate to the theory, that some users and producers of spatial data may feel that visualization of uncertainty reduces the perceived quality of data by end users (Kinkeldey & Schiewe 2014). This is directly addressed by Question 6, linked to the potential usability of Uview as a visualization tool for end users. Most evaluators felt that visualization would not degrade the perception of end users regarding data uncertainty. As visualization is based on accuracy assessments that are already provided, the quality of the data should already be known. The *Graduates* group members felt that visualization would not degrade the perception of GIS professionals, as they already understand the inherent nature of uncertainty within geospatial datasets. *Graduates* members were, however, concerned that it may affect how end users (non-professional) appreciate the data, because they may assume complete accuracy of the data without inspecting the metadata. In this case, a visualization would highlight uncertainty in the geospatial data of which they were previously unaware. One *Industry* group member, with much experience of uncertainty visualization, felt that visualization aided positively in the communication of uncertainty. Although end users may have a negative perception at first, when the concept of uncertainty and uncertainty visualization is explained to them, they generally had a positive response to it. Evaluators from the *Industry* group also indicated that it can serve to enhance the data as an explicit quality assurance given to the client adding value to the work done.

Therefore, according to the evaluators, visualization of uncertainty does not pose a major risk of compromising the end user perception of data quality, but it may require some additional description of the nature of spatial data and associated uncertainty. It also needs to be appreciated that uncertainty does not mean that data quality is poor; in the long term

educating end users (clients) of this will improve the understanding of the nature of spatial data and the perceived quality of that data. Although there may be risks that data is not perceived as of good quality if a visualization of uncertainty is given, these can be dealt with through educating clients and developing better uncertainty management methods. Visualization and the new understanding of uncertainty may also lead to a perception of an open and transparent model of supplying data, which aids in building trust between industries and clients.

6.1.2.3 Management of uncertainty

This question was dealt with in two phases. Firstly, it was established how uncertainty is currently managed in the geospatial data processing workflow, and if visualization could aid in such management. This was addressed through Questions 2 and 4, which served to establish the conceptual need for a visualization software tool. Then, the response to Question 5 sought to establish how communication of uncertainty to end users took place and whether a visualization tool could, at least in theory, be useful. Question 2 was posed before demonstration of Uview, while Questions 4 and 5 followed after the evaluators had been introduced to the software tool. All evaluators agreed that there is no standard method of managing uncertainty in geospatial data; uncertainty is dealt with in a pragmatic manner by considering each job in terms of its requirements, the costs involved and each client's unique requests. In some cases, there is no management, and the data is simply used as supplied, while in other cases, there is a conscious effort to ensure high data quality and its effective communication.

One of the most interesting comments mentioned by a respondent from *Industry* group was: "the poorer data quality is, the more prominent the warning about the use and quality of the data is in reports." This approach was especially used in exploratory research and data creation, where good accuracy assessments are not always available. With better data quality, the warnings become less prominent, relegated to an accuracy assessment section to be consulted specifically by the user to obtain the statistics. Only two evaluators, one from the *Industry* group and one from the *Academic* group, referred to currently using visualization as a method of data quality management, both to show where the data quality may be poor, as well as to use it for targeted correction of data they develop. These visualizations would then be communicated via a report or article. All evaluators agreed that visualization with a tool such as Uview could be a powerful mechanism for managing data quality. They unanimously

agreed that Uview is simple to use and can easily be incorporated into a workflow to check and communicate data quality.

Meanwhile, these results from the questionnaires (in Chapter 3) support the theory that a standardised methodology for dealing with uncertainty in geospatial datasets (Tegtmeier et al. 2007) does not exist.

6.1.2.4 Suggestions for software tool

Finally, Question 7 set out to determine what suggestions or remarks each evaluator had for Uview, be it improvements, additional functionality or shortcomings that may need to be critically addressed.

The most frequent suggestion was that the uncertainty metrics used by Uview should be explained in more detail. This was addressed by updating Uview documentation, in order to give information on the uncertainty statistics and how they are calculated in an 'About page' on the Uview interface.

One *Industry* group member suggested that the way Uview performs point sampling should be changed; Uview currently samples at the location of the coordinates of the supplied points (see Figure 6.2).

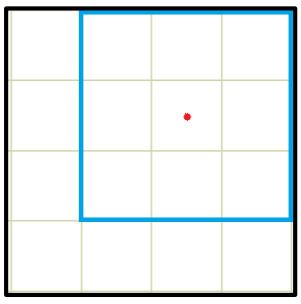


Figure 6.2 Sampling grid

Currently, the sample value is compared to the reference value contained in the shapefile or on the reference raster. As the supplied point may not fall in the centre of the sampled pixel, it was suggested to rather sample a 3 X 3 neighbourhood (blue square in Figure 6.2) of pixels around the point and using the mean value for the value at the point.

An *Academic* group member who uses visualizations on a regular basis, felt that the symbology (class colours for uncertainty) could be improved to show greater difference between classes. An *Industry* group evaluator suggested that Uview should not only visualize absolute difference, but also true difference, which would show the magnitude as well as the

direction of the difference. It is important to note here that, although it is not a built-in visualization option, the true values are calculated and provided as an attribute in the visualization shapefile. The user may visualize the attribute data in any suitable manner.

A possible shortcoming pointed out by the two *Academics* group members is the question of scale or resolution of the uncertainty visualization. The best resolution result Uview can provide is based on the maximum distance between reference points. A reasonable sample of reference points (more than 30) is also needed to provide a useful visualization. Another *Industry* group evaluator suggested that Uview should be improved to include an internal evaluation of data quality. This may entail scanning the dataset to highlight areas where rapid change between cell values occurs, indicating potentially a processing error, faults in the data indicated by 'nodata' (NULL) values, or internal inconsistencies. Finally, there was a request for Uview to also cater more comprehensively for discrete data uncertainty visualization and incorporate more statistics.

Overall it was found that Uview can be a useful product. It is easy to use and provides useful statistics and visualizations. When using Uview, users should understand the implications of the statistics. Some of the more important points to note and address in future releases are:

- no internal validation of datasets;
- scale issues;
- visualization contrast;
- visualizing negative and positive differences.

A use for Uview suggested by both *Academics* and *Industry* groups members, which was not considered during the design of Uview, is to perform change analysis for raster datasets. The statistics produced by Uview can be used to show areas of change between two continuous raster datasets as long as enough sampling points are used. When comparing two raster datasets, a very fine point grid can be used for a good resolution result. This is a potential use that anyone using Uview can exploit, after reading the 'About page' and being comfortable that it is using the correct statistics that the user understands and requires.

As a tool for accuracy assessment and communication of uncertainty, Uview is a success as it is easy to use, easily incorporated into workflow and provides useful visualization communicating the quality of a dataset, especially in conjunction with the statistics also provided.

6.2 CHAPTER FINDINGS

Uview provides both the statistics used in a traditional accuracy assessment (of continuous raster data) as well as a visualization to communicate the spatial element of data quality, which is the unique element that distinguishes geospatial data from any other form of data. The evaluators could define uncertainty, but had no uniform way of dealing with uncertainty in geospatial data, communicating on matters of data quality and uncertainty, or of managing it. Some companies may not even have a good internal communication of data quality as indicated by one *Graduate* member. Perhaps the most pragmatic method of uncertainty management mentioned, was that the lower the data quality, the more prominent the warnings about product quality; the better the quality the less prominent the warnings.

All evaluators interviewed responded positively to Uview; all agreed that visualization, together with statistics, could improve the understanding of the data quality. This is in agreement with the findings of Perer & Shneiderman (2009), that visualization together with statistics can improve understanding and aid in faster analysis. Though there is some concern amongst *Graduates* members that visualization may depreciate the perceived quality of data by end users, if uncertainty and data quality are explained to users in a meaningful way, this could improve their understanding of both the data and the quality thereof.

Whilst there are some improvements that can be made to Uview, it is already a useful tool that can effectively communicate uncertainty. Currently, Uview is limited to continuous raster data and requires reference data, whilst the quality of the Uview products (statistics and visualization) is also largely related to the quality of the provided reference dataset (both problems also related to any accuracy assessment). Uview has present limitations, but as long as they are understood, Uview can be used as a successful accuracy assessment and visual communication tool.

CHAPTER 7 REVIEW OF RESEARCH

This research has set out to evaluate the understanding of data quality and propose a useful method of visualizing uncertainty in geospatial data. The following research question was addressed:

- Can uncertainty which is inherent in spatial data or produced through modelling be visualized to facilitate understanding of the data quality by those who use and produce spatial data?
- How does visualization of uncertainty affect users' perception of spatial products produced?

To address these questions, a set of five objectives was established and investigated. These objectives are:

1. Establish a baseline perception of general data quality amongst users and producers when working with spatial data;
2. Evaluate available uncertainty visualization tools for raster data through literature;
3. Develop software tool for uncertainty visualization of continuous raster data;
4. Generate visualization scenarios to test the software tool;
5. Compare the effect of statistical and visualized uncertainty on the perception of users and producers of spatial data, as well as on their decision making.

Three tasks were set to enable the establishment of a working model for this research, split between Chapter 3 to Chapter 6. Task 1 was to evaluate the South African perception on data quality; Task 2 was to develop and build an uncertainty visualization tool, based on the findings of literature as well as on the results of Task 1; lastly Task 3 entailed evaluating the tool through modelling and focussed interviews.

These three tasks together with the literature review, achieved the aim of determining to what extent users of spatial data understand the quality of their data, successfully covered the development of a tool that can enhance the communication of the overall quality of datasets.

7.1 TASK 1

The first task was achieved through a two-pronged approach. International literature was reviewed, together with regulations from the South African Geomatics Council (previously

South African Council for Professional and Technical Surveyors (PLATO)) for registration as a geographic information science (GISc) professional. This then lead to the development of a questionnaire. The main themes of the questionnaire were:

- awareness of the uncertainty inherent in geospatial data;
- how uncertainty is dealt with in the current workflow;
- how well geospatial professionals understand data quality when consulting the accuracy assessment;
- which visualization from literature was the easiest to understand;
- the value of visualization as a representation of uncertainty.

The results from respondents of the survey indicated unanimously, that uncertainty is known by nearly all those working with geospatial data. This was a successful start, but when it came to how uncertainty is managed, as indicated by the research of Kinkeldey & Schiewe (2014) and Tegtmeier et al. (2007), it was found that uncertainty is not dealt with in a uniform way and varies from person to person. As Knight (1921) in Foss & Klein (2012) holds, uncertainty is dealt with better by some individuals than others. The findings of the survey revealed three categories for dealing with uncertainty; 1) trying to improve the data and communicating the data quality (uncertainty) to those who will use it; 2) using data on a fit for purpose basis; 3) ignoring the uncertainty of the dataset. This inadequacy with uncertainty management became even more problematic, when it was indicated that only about 60% of respondents asked to see an accuracy report on data they use. Moreover, only just short of 50% of those who indicated that they try to improve data and also communicate the uncertainty in the data they work with, do not ask for an accuracy assessment of the data. This raises questions as to: a) how the extent of data quality can be gauged and improved upon when the extent of its accuracy is unknown; and b) how data can be evaluated as fit for a purpose when the quality of that data is not known? Some respondents also indicated that they assume a high data quality just because the data was supplied by one or another supposedly reliable body.

The visualization scheme developed by Howard & MacEachren (1996) for their tool R-Vis, was indicated by respondents as the easiest visualization to understand as it puts data into distinctly differently shaded classes of certainty. The visualization can be seen in Chapter 3 (Image 5 of Figure 3.4). When asked whether the respondents would like a visualization of

uncertainty, 40% of those that ask for an accuracy assessment indicated they would like a visualization, whilst 70% of those that do not look at the accuracy assessment would like a visualization. This is in agreement with Tegtmeier et al. (2007), who found that some professionals may not want a visualization as it may devalue the perception of the quality of their data. Nevertheless, when one considers the large group that does not ask for accuracy assessments, together with the large group that would like a visualization, visualization of uncertainty can bring the much-needed knowledge of a dataset's quality to a wider audience. The lack in understanding of uncertainty and ignorance of some about the quality of data they use, can perhaps be linked to the short time allocated in the academic model suggested by PLATO before registration. Less than four percent of the required teaching time is spent on uncertainty teaching.

7.2 TASK 2

The aim of this task was to develop and build an uncertainty visualization tool. The requirements for the tool were largely informed by the findings from Task 1, where it was found that there is a need for a visualization tool which can increase the understanding of data quality. The first endeavour was to review some of the available tools, out of which four were chosen: R-Vis by Howard & MacEachern (1996); UncertWeb an as yet unpublished web based visualization tool funded by the European Commission (EC); Aguila the visualization tool of the tool PCRaster; and lastly UVIS by Alberti (2013), a tool that uses type A probabilistic methods for visualization (Gerharz et al. 2012; Karssenberg et al. 2010).

Of the four tools evaluated, only Aguila is currently available to be used. Aguila was however difficult to install (on a windows machine) and the learning curve to use it is particularly steep. The only tool that was easy to use albeit with pre-prepared data (a demonstration model only), was UVIS. Development of UncertWeb stopped when funding discontinued in 2013, with no discernible product yet available. The main requirements for Uview (the tool developed as a result of this study) were:

- freely available;
- easy to install;
- easy to use;
- provides statistics;
- visualizations with colour blind user support.

To achieve these requirements, QGIS and in particular a QGIS plugin, was chosen as the platform. QGIS provides cross-platform support, easy installation and is also freely available to all. It also satisfies the requirement of Kinkeldey & Schiewe (2014), who found that users would like a plugin for the common software they use (QGIS or ArcMap), to easily incorporate into their workflow. With a minimum of two and a maximum of three inputs needed, Uview is easy to use and no complex input commands are needed. Uview is not only a visualization tool but also a statistical accuracy assessment tool. The visualizations are based on two statistics: z-score and modified z-score. The z-score and the modified z-score both measure the distribution from the mean of individual reference points. An absolute difference visualization is also an option to provide the raw difference between reference data and dataset being evaluated. Further, an overall visualization combining all three z-score based measures is available to provide a general statistical overview visualization. Together with the visualizations within the output shapefile attribute table, the standard deviation, mean absolute error (MAE) and root mean square error (RMSE) are also provided. This thus provides data for traditional accuracy assessments, as well as the visualization to communicate the spatial nature of the uncertainty. Visualizations are provided for colour blind as well as normal vision users, in order to enable all users to be able to extract maximum use out of them. Using a shapefile product for evaluating raster data achieves two major advantages. Firstly, the original raster dataset is not modified, secondly the user of the tool is free to change the symbology of the visualization in QGIS in any way that suits them. Once the tool was created, it was then reviewed to test both the value and the usefulness of it.

7.3 TASK 3

Task 3 was to evaluate the software tool through statistics, simulation and qualitative interviews. The description of the task was divided between Chapter 5 and Chapter 6, with Chapter 5 introducing the tool, the installation and functionality as well as the simulation and statistical evaluation. Chapter 6 was an evaluation of the tools usability, and enhancement of accuracy assessments with the visual aspect, through the interviewing of 12 individuals involved in geospatial data.

7.3.1 Evaluation one

To evaluate Uview, a multipronged approach was used. Firstly, four watershed products from two different areas were compared with the Uview visualization for the original digital elevation models (DEMs). These watersheds were derived from two high quality reference

DEMs (Ref A and Ref B), and one from a similar resolution DEM (Test A and Test B). One of the DEMs, Test A, was partially corrected at the three highest categories of uncertainty as indicated by Uview to match the Ref A. A watershed product was then developed from this partially corrected DEM (Test A-Cor) and evaluated against the Uview visualization and the other two watershed products (Basin-RA and Basin-TA). Following this, a Monte Carlo based simulation (Zandbergen 2011) was run on Test A using the statistics derived from Uview, in order to view where higher and lower probability watershed boundaries occur. This product was then once again compared to the Uview uncertainty visualization and the Basin-RA.

The two study areas were compared using the naming convention of Test A and Test B, as the data to be evaluated, and Ref A and Ref B were the higher quality reference data used for the accuracy assessment of Test A and Test B. All visualizations from Uview provided a similar result flagging similar areas, as can be seen in Chapter 5. The modified z-score best flagged outliers and extreme values. The normal z-score and the absolute values z-score provided highly related results with the Overall Visualisation Index (OVI), providing an average between the modified z-score and normal z-score. All the Uview visualizations confirmed with Zandbergen (2011) and Weng (2012) that uncertainty clusters in specific areas. Investigating the relationship between elevation, slope and topographic terrain ruggedness index and uncertainty, the strongest correlation was found between topographic terrain ruggedness and uncertainty. It was found in both study areas, that above a topographic index of four, the probability of finding outliers (more uncertainty) becomes close to 50%. Elevation and slope on the other hand showed a weak correlation with uncertainty, although there was slightly higher correlation between slope and uncertainty. All of these tests however indicated that uncertainty can occur at any location but higher uncertainty tends to cluster.

When comparing the Uview products with the watershed products, it was found that uncertainty at one point may not always cause a difference at that point, however this uncertainty may affect products at another location away from the original uncertain area in the original dataset. Test A was partially corrected to match Ref A, based on the three highest classes of uncertainty as indicated by Uview, to create dataset Test A-Cor and run through the watershed model. The result provided a watershed model that deviated from Basin-RA and Basin-TA at some locations, but at other locations it related to one or the other dataset. This illustrated that at key places, even small deviations may create differences to the resulting

product of a dataset, and even a partially corrected dataset may not provide the results of a higher quality dataset.

When comparing the probability model basin in Chapter 5 with the delineation Basin-TA, it was shown that lower probability does not always link up to areas where Uview flagged high uncertainty, or where Basin-RA and Basin-TA deviate. The conclusions to be drawn from these results together with the Test A-Cor results, indicate that, although Uview can produce a visual aspect of the quality spatially and visually, it is necessary to really know for what the input dataset will be used. It is also just as important to understand how the processing of the dataset works, and what elements of a dataset the product of this processing is sensitive to, as large deviations in some places may not cause problems for a derived product, but small deviations in key areas may have large consequences for a derived product.

The conclusion that can be drawn from this analysis is that Uview does provide a usable result. Uview can be used to indicate the spatial nature of uncertainty through the visualizations, and the statistics are comparable to traditional accuracy assessment outputs. Furthermore, Uview products can be used to investigate new relationships such as those between topographic terrain ruggedness index and uncertainty in DEMs.

7.3.2 Evaluation two

Chapter 6 focussed on the qualitative interviews aspect of Task 3. For this investigation three questions were posed: 1) how does Uview add value to the understanding of statistics, as indicated by Perer & Shneiderman (2009)?; 2) Is there merit in the belief by producers that a visualization may reduce the perceived quality of a dataset (Kinkeldey & Schiewe 2014)?; 3) Is there a uniform method which users and producers apply to manage uncertainty (Tegtmeier et al. 2007)?

In-depth interviews with twelve people who regularly work with geospatial data were conducted to answer these questions. After an introduction to traditional accuracy assessment statistics of a selected dataset, a demonstration of Uview was given. Evaluators uniformly agreed that the visualization from Uview provides a better understanding of the uncertainty statistics, as well as the spatiality of the uncertainty. Uview was found to be easy to use, and can be used to evaluate and improve data whilst creating a dataset (producer), as well as packaged with supplied data to give an intuitive quality assurance to the client on the data standard. This demonstrates that visualization can contribute to the understanding of statistics as indicated by Perer & Shneiderman (2009).

Contrary to the findings of Kinkeldey & Schiewe (2014), most respondents agreed that visualization of uncertainty will not devalue the perception of a dataset quality, but rather enhance the understanding of the data. It was however commented, that visualization can affect the perception of the data for those with no understanding of geospatial data, but would not to professionals and those with experience and understanding of geospatial data. Sufficient training on data quality and visualization to end users by data distributors may solve this and visualization can lead to a better understanding of uncertainty by all.

When asked about knowledge and management of uncertainty, evaluators confirmed the findings in Chapter 3, that whilst all were aware of uncertainty, there were different methods of dealing with it. Two evaluators mentioned visualizing uncertainty, however most stated documentation or metadata as a description of data quality. On the other hand, three interviewees noted that they used whatever data is supplied without care to the quality. The findings of Tegtmeier et al. (2007), that there is no uniform method of dealing with uncertainty, is thus echoed here.

7.3.3 Suggestions for Uview

With the overall finding that Uview is an easy-to-use tool that produces valuable results, evaluator recommendations and suggestions covered: i) a slightly revised sampling technique involving a 3x3 sampling grid; ii) a warning or minimum requirement statement about the number of points needed to provide a worthwhile visualization; iii) a visualization that groups positive and negative differences separately; iv) greater contrast between visualization classes; v) include enhancement to evaluate within one dataset without reference data, and flag inconsistent cells that have no data or anomalies within the dataset; vi) better description of the uncertainty statistics used (now provided); and vii) use Uview for change-over-time analysis, such as climate change when comparing two raster datasets from two time periods.

In terms of the number of points needed to provide a worthwhile visualization, the best visualization resolution given is the maximum distance between two points. One academic however, indicated that this is a problem with accuracy assessments in general, and that users should just follow the same requirements as for any accuracy assessment.

In the different grouping of positive and negative differences with greater contrast, Uview provides all values in the visualization shapefile's attribute table, providing users with freedom to change the symbology as they choose, as well as to query all values and statistics

at any point. Meanwhile, a change-over-time analysis could provide useful results, is feasible and could be implemented by users of the tool.

7.3.4 Task 3 view

Overall the findings from this task were that a professional working with geospatial data must not only understand the quality of data and the spatial extent of uncertainty, but also have a good comprehension of geoprocessing tools and the effect of processing on the data. Only when both the quality of the data, as well as the sensitivity of the data to the specific processing steps are considered, can an educated decision be taken on the fitness of the data for use.

Further, despite its shortcomings, evaluators agreed that Uview was easy to install and understand, and provided valuable information for comprehending data quality statistics. Uview was found to meet all five its requirements, to be: 1) freely available; 2) easy to install; 3) easy to use; 4) providing statistics; 5) provide visualizations with colour blind user support.

7.4 LIMITATIONS

Though this study has contributed to both the understanding of uncertainty in South Africa as well as the visualization of uncertainty, there are some limitations. The number of respondents for both questionnaire and focussed interviews was small. This may be as a result of the size of the geospatial community, as well as the recruitment instruments. However, the results cannot be ignored, as these numbers are in line with the respondents from other international studies of similar content. Uview itself has limits in that the resolution of the output is dependent on the input reference data, but also that the statistics used to visualize uncertainty can only compare the results within the dataset itself with reference to its mean. There is no statistic incorporated yet that can compare the dataset with an idealized perfect dataset. In addition, Uview focusses on continuous raster data only.

7.5 RECOMMENDATIONS FOR FURTHER RESEARCH

Each of the three tasks in this study could be further developed into standalone topics. This research has established a baseline for uncertainty perception, which can be expanded to a wider audience. The perception of South African geospatial data users and producers needs to be further investigated, as well as the reasons behind why they manage data quality the way that they do currently. For this, a geospatial data user or producer must clearly be defined.

Furthermore, though knowledge of data quality is included in the PLATO model, there is not a strong enough emphasis on how data quality should be managed. This section deserves research, with tangible recommendations on how to improve the regulations in its own right and not only as a part of a larger study.

Geostatistics, especially those that relate to the spatial nature of data and the quality thereof, also needs further research. Such research does not necessarily need to culminate in the production of tool, but rather in methods to compare a dataset with an idealized perfect dataset and how to best communicate the spatial nature of data uncertainty. This is critical as the current accuracy assessments provide a global statistic on spatial data that inherently varies over the distance of the dataset locally.

GIS in itself is in a process of change. The digital age is moving towards online productivity and online applications. Analysis and processing of data is becoming an online process, thus for the next generation of uncertainty visualization tools, cloud based options should be considered.

7.6 SUMMARY OF RESEARCH RESULTS

Uncertainty is a constant, in humans and in geospatial data. The philosopher René Descartes has said the only thing humans can be sure of is that each individual in themselves exists (Descartes 1951). Uncertainty is as much part of the human condition as it is part of any geospatial data. In the global and academic context, much is known about uncertainty with many studies on the statistics and some into visualization thereof. However, although much is known, those using geospatial data are still not always as aware of the uncertainty in their data as they should be. Uncertainty is often not well documented or understood. The knowledge is fractured and although nearly all know of the presence of uncertainty, very few are aware of what the implications may be or are even aware of the quality of their data.

Some South African users and producers of geospatial data do not always consult the accuracy assessment of data they use, nor is there a uniform way of managing uncertainty. As such, the hypothesis in Chapter 1, that users and producers of spatial data are conscious of the quality of their data can be rejected. Visualization could bring data uncertainty to the attention of a wider audience, particularly to those that currently do not look at accuracy assessments. With Uview, statistics are provided in combination with a visualization and can be used as a full accuracy assessment tool for continuous raster data, rather than just a visualization tool. Visualization interwoven with traditional statistical accuracy assessments

can improve the understanding of geospatial data quality and its inherent uncertainty. Better education around data quality and uncertainty, as well as effective uncertainty management tools or workflows may further lead to more responsible users of geospatial data. This can create a situation where everyone is careful and knowledgeable of the data they use as well as the processes they use it for.

REFERENCES

Aguilar F, Agüera F & Aguilar M 2007. A theoretical approach to modeling the accuracy assessment of digital elevation models. *Photogrammetric Engineering & Remote Sensing* 73, 12: 1367-1379.

Aguirre-Gutiérrez J, Carvalheiro LG, Polce C, Van Loon EE, Raes N, Reemer M & Biesmeijer JC 2013. Fit-for-purpose: Species distribution model performance depends on evaluation criteria – Dutch hoverflies as a case Study. *PLOS ONE* 8, 5.

Alberti K 2013. Web-based visualization of uncertain spatio-temporal data. Master's thesis. Utrecht: Universiteit Utrecht.

American Academy of Ophthalmology 2014. Caucasian boys show highest prevalence of color blindness among preschoolers [online]. Available from: http://www.aao.org/newsroom/release/color-blindness-among-preschoolers-ophthalmology-journal-study.cfm [Accessed 14 March 2015].

Asaka S 2016. A primer to social theory: Towards understanding theory construction and application in sociological discourse. *Journal of Humanities and Social Science* 21, 5.

Baruch Y & Holtom BC 2008. Survey response rate levels and trends in organizational research. Human Relations 61, 8: 1139-1160.

Bastin L, Cornford D, Jones R, Heuvelink GBM, Pebesma E, Stasch C, Nativi S, Mazzetti P & Williams M 2013. Managing uncertainty in integrated environmental modelling: The UncertWeb framework. *Environmental Modelling & Software* 39: 116-134.

Beaulieu ND & Epstein AM 2002. National committee on quality assurance health-plan accreditation: Predictors, correlates of performance, and market impact. *Medical Care* 40, 4: 325–337.

BIPM 2008. *Guide to the Expression of Uncertainty in Measurement*. Geneva: International Organization for Standardization.

Bishop D & Karadaglis C 1996. Combining GIS-based environmental modelling and visualization: another window on the modeling process. In: *Third International Conference/Workshop on Integrating GIS and Environmental Modeling*. The National Center for Geographic Information and Analysis: Buffalo.

Blackstone A 2016. *Principles of sociological inquiry*. Washington: Flat World Education.

Bordoloi UD, Kao DL & Shen HW 2004. Visualization and exploration of spatial probability density functions: A clustering-based approach. *Visualization and Data Analysis* 57.

Bostrom A, Anselin L & Farris J 2008. Visualizing seismic risk and uncertainty. *Annals of the New York Academy of Sciences* 1128, 1: 29-40.

Brewer CA, MacEachren AM, Pickle LW & Herrmann D 1997. Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers* 87, 3: 411-438.

Buckley DJ 1997. *The GIS primer: An introduction to Geographic Information Systems*. Fort Collins: Pacific Meridian Resources.

Burg MB, Peeters H & Lovis WA (Eds) 2016. *Uncertainty and sensitivity analysis in archaeological computational modelling*. Switzerland: Springer International Publishing.

Candes EJ, Romberg JK & Tao T 2006. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics* 59, 8: 1207-1223.

Chai T & Draxler RR 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* 7, 3: 1247-1250.

Chai T, Kim HC, Lee P, Tong D, Pan L, Tang Y, Huang J, McQueen J, Tsidulko M & Stajner I 2013. Evaluation of the United States national air quality forecast capability experimental real-time predictions in 2010 using air quality system ozone and NO2 measurements. *Geoscientific Model Development* 6, 2: 1831-1850.

Chi EH 2000. A taxonomy of visualization techniques using the data state reference model. in *IEEE Symposium on Information Visualization*. IEEE: New York.

Congalton RG & Green K 2008. *Assessing the accuracy of remotely sensed data: Principles and practices*. 2nd ed. Taylor & Francis: Boca Raton.

Congalton RG 1997. Exploring and evaluating the consequences of vector-to-raster and raster-to-vector conversion. *Photogrammetric Engineering & Remote Sensing* 63, 4: 425-434.

Conger S 2004. *A review of colour and cartography in avalanche danger visualization*. Columbia: University of British Columbia.

Conolly J & Lake M 2006. *Geographical information systems in archaeology*. Cambridge: Cambridge University Press.

Couclelis H 2003. The certainty of uncertainty: GIS and the limits of geographic knowledge. *Transactions in GIS* 7: 165–175

Davis B 2001. *GIS: A Visual Approach*. Albany, New York: Delmar Thomson Learning.

De Gennaro M, Paffumi E, Scholz H & Martini G 2014. GIS-driven analysis of e-mobility in urban areas: An evaluation of the impact on the electric energy grid. *Applied Energy* 124: 94-116.

De Graaff N 2013. Communicating uncertainty in spatial data: Exploring the usefulness in municipal information chains. Master's thesis. Amsterdam: University Amsterdam, Faculty of Earth and Life Sciences.

Del Campo AG 2012. GIS in environmental assessment: A review of current issues and future needs. *Journal of Environmental Assessment Policy and Management* 14, 01.

Department of Public Service & Administration 2008. *Policy on free and open source software use for South African government*. Pretoria: Department of Public Service and Administration.

Descartes R 1951. *A discourse on method*. New York: Dutton.

DiBiase D, DeMers M, Johnson A, Kemp K, Luck AT, Plewe B & Wentz E (eds) 2006. *GI S&T body of knowledge: University consortium for geographic information science*. Penn State University: Association of American Geographers.

Dol W & Verhoog D 2010. *Small to medium-scale focused research project under the seventh framework program*. Bonn: University of Bonn.

Du DZ & Hwang F 1992. *Computing in euclidean geometry*. Singapore: World Scientific.

Du Plessis H & Van Niekerk A 2012. A curriculum framework for Geographical Information Science (GISc) training at South African universities. *South African Journal of Higher Education* 26, 2: 329-345.

Du Plessis H & Van Niekerk A 2014. A new GISc framework and competency set for curricula development at South African universities. *South African Journal of Gemomatics* 3, 1: 1-12.

Du Plessis HJ 2015. A methodology for assessing geographical information science professionals and programmes in South Africa. Stellenbosch: Stellenbosch University, Department of Geography and Environmental Studies.

Edelsbrunner H 2014. *A short course in computational geometry and topology*. Cham: Springer.

Elo S & Kyngäs H 2008. The qualitative content analysis process. *Journal of Advanced Nursing* 62 1: 107-115.

ESRI 2016. FAQ: What is the difference between the basin and watershed tools from the spatial analyst toolbox? [online]. Available from: http://support.esri.com/technical-article/000012352 [Accessed 29 September 2016].

ESRI s.a.a. *DEM | Definition - Esri Support GIS Dictionary* [online]. Available from: http://support.esri.com/other-resources/gis-dictionary/search/DEM [Accessed 15 April 2015].

ESRI s.a.b. *vector | Definition - Esri Support GIS Dictionary* [online]. Available from: http://support.esri.com/other-resources/gis-dictionary/search/vector [Accessed 15 April 2015].

ESRI s.a.c. *vector data model | Definition - Esri Support GIS Dictionary* [online]. Available from: http://support.esri.com/sitecore/content/support/Home/other-resources/gis-dictionary/term/vector%20data%20model [Accessed 15 April 2015].

ESRI s.a.d. *raster | Definition - Esri Support GIS Dictionary* [online]. Available from: http://support.esri.com/other-resources/gis-dictionary/search/raster [Accessed 15 April 2015].

ESRI s.a.e. *standard deviation | Definition - Esri Support GIS Dictionary* [online]. Available from: http://support.esri.com/other-resources/gis-dictionary/search/standard%20deviation [Accessed 15 April 2015].

SeoESRI s.a.f. *z-score | Definition - Esri Support GIS Dictionary* [online]. Available from: http://support.esri.com/other-resources/gis-dictionary/search/z-score [Accessed 15 April 2015].

ESRI s.a.g. *slope | Definition - Esri Support GIS Dictionary* [online]. Available from: http://support.esri.com/other-resources/gis-dictionary/search/slope [Accessed 15 April 2015].

Fagan C & Maidment DR s.a. *Converting a LAS data to a DEM and performing a watershed delineation*. University of Texas: Austin.

Feizizadeh B, Jankowski P & Blaschke T 2014. A GIS based spatially-explicit sensitivity and uncertainty analysis approach for multi-criteria decision analysis. *Computers & Geosciences* 64: 81-95.

Fisher PF (ed) 1995. *Innovations in GIS 2*. Taylor & Francis: London.

Fisher PF 1992. First experiments in viewshed uncertainty: The accuracy of the viewshed area. *Photogrammetric Engineering and Remote Sensing* 58, 3: 345-352.

Foody GM & Atkinson PM (eds) 2003. *Uncertainty in remote sensing and GIS.* Chichester: John Wiley and Sons.

Foody GM 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment* 80, 1: 185-201.

Foss NJ & Klein PG 2012. *Organizing entrepreneurial judgment: A new approach to the firm*. Cambridge: Cambridge University Press.

Fowler JMS 2011. Cartographic communication of point level uncertainty. Master's thesis. Columbia: University of South Carolina.

Frick PJ, Barry CT & Kamphaus RW 2009. *Clinical assessment of child and adolescent personality and behavior.* New York: Springer.

Friederich 2014. Comparison of ArcGIS and QGIS for applications in sustainable spatial planning. Master's thesis. Vienna: Universität Wien.

FUGRO s.a. *LiDAR mapping fact sheet*. Cape Town: FUGRO.

GeoApt LLC s.a. QGIS Plugin Builder — QGIS Plugin Builder 2.0 documentation [online]. Available from: http://geoapt.net/pluginbuilder/ [Accessed 8 June 2016].

Gerharz L, Pebesma E & Hecking H 2010. Visualizing uncertainty in spatio-temporal data. In: N.J. Tate and P.F. Fisher (eds), *Proceedings of the ninth international symposium on spatial accuracy assessment in natural resources and environmental sciences*, 20–23 July. Leicester.

Gerharz L, Senaratne H, Autermann C, Truong NP, Heuvelink GBM, Williams M, Pebesma E, Stasch C & Cornford D 2012. *Tools for communicating and visualising uncertainties*. Aston: Aston University.

Goodchild MF 1996. *GIS and environmental modelling: Progress and research issues*. GIS World Books: Fort Collins.

Google Earth s.a. Google Earth [online] Available from: https://www.google.com/earth/ [Accessed 31 October 2015].

Google Maps s.a. Google Maps [online]. Available from: https://www.google.co.za/maps [Accessed 16 July 2016].

Habib A & Van Rens J 2008. *Quality assurance and quality control of LiDAR systems and derived data*. Bethesda: American Society for Photogrammetry & Remote Sensing.

Harris H & Jarvis C 2011. *Statistics for geography and environmental science*. New York: Taylor & Francis.

Hessler J 2014. History of GIS and Early Computer Cartography Project. Redlands: ESRI 17-20.

Hirano A, Welch R & Lang H 2003. Mapping from ASTER stereo image data: DEM validation and accuracy assessment. *ISPRS Journal of Photogrammetry and Remote Sensing* 57, 5-6: 356-370.

Howard D & MacEachren AM 1996. Interface design for geographic visualization: Tools for representing reliability. *Cartography and Geographic Information Systems* 23, 2: 59-77.

Huggel C, Schneider D, Miranda P, Delgado Granados H & Kääb A 2008. Evaluation of ASTER and SRTM DEM data for lahar modeling: A case study on lahars from Popocatépetl Volcano, Mexico. *Journal of Volcanology and Geothermal Research* 170, 1-2: 99-110.

Hunsaker CT, Goodchild MF, Friedl MA & Case TJ (eds) 2001. *Spatial uncertainty in ecology*. Springer: New York.

Iglewicz B & Hoaglin DC 1993. *Volume 16: How to detect and handle outliers*. Wisconsin: ASQC Quality Press.

Jacquez GM 2012. A research agenda: Does geocoding positional error matter in health GIS studies?. *Spatial and Spatio-temporal Epidemiology* 3, 1: 7-16.

Jefferies D & Evetts J 2000. Approaches to the international recognition of professional qualifications in engineering and the sciences. *European Journal of Engineering Education* 25, 1: 99-107.

Jenny B & Hurni L 2011. Studying cartographic heritage: Analysis and visualization of geometric distortions. *Computers & Graphics* 35, 2: 402-411.

Jenny B & Kelso NV 2007. Color design for the color vision impaired. *Cartographic Perspectives* 58: 61-67.

Jiang B 2012. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer* 65, 3: 482-494.

Kaplan RM & Saccuzzo DP 2008. *Psychological testing: Principles, applications, and issues*. 7th ed. Belmont: Wadsworth Publishing.

Karssenberg D, Schmitz O, Salamon P, De Jong K & Bierkens MFP 2010. A software framework for construction of process-based stochastic spatio-temporal models and data assimilation. *Environmental Modelling & Software* 25, 4: 489–502.

Kaye NR, Hartley A & Hemming D 2012. Mapping the climate: guidance on appropriate techniques to map climate variables and their uncertainty. *Geoscientific Model Development* 5, 1: 245-256.

Kinkeldey C & Schiewe J 2014. Expert interviews about the use of visually depicted uncertainty for analysis of remotely sensed land cover change. Hamburg: HafenCity University Hamburg.

Kinkeldey C, MacEachren A, Riveiro M & Schiewe J 2015. Evaluating the effect of visually represented geodata uncertainty on decision-making: systematic review, lessons learned, and recommendations. *Cartography and Geographic Information Science* 1-21.

Kinkeldey C, MacEachren AM & Schiewe J 2014. How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal* 51, 4: 372-386.

Kraak MJ & Ormelling FJ 2011. *Cartography visualization of spatial data*. New York: Guildford Press.

Krygier J & Wood D 2005. *Making maps*. Guilford Press: New York.

Krygier J 2014. Geography 353 [online]. Available from: http://krygier.owu.edu/krygier_html/geog_353/geog_353_lo/geog_353_lo09.html [Accessed 24 February 2015].

Lacroix V 2009. Raster-to-vector conversion: Problems and tools towards a solution a map segmentation application. In: *ICAPR '09 Proceedings of the 2009 Seventh International Conference on Advances in Pattern Recognition* 318-321.

Lawhead J 2015. *QGIS python programming cookbook*. Birmingham: Packts Publishing Ltd.

Lee B 2009. Spatial pattern of uncertainties: An accuracy assessment of the TIGER files. *Journal of Geography and Geology* 1, 2.

Liang X, Kang A & Pettorelli N 2016. Understanding habitat selection of the vulnerable wild yak Bos mutus on the Tibetan plateau. *Oryx* 1-9.

Longley PA, Goodchild MF, Maguire DJ & Rhind DW (eds) 1999. *Geographical Information Systems, Volume 1*. John Wiley & Sons, Inc: San Francisco.

Longley PA, Goodchild MF, Maguire DJ & Rhind DW 2005. *Geographical information systems and science*. 2nd ed. Chichester: John Wiley & Sons Ltd.

Lucieer A 2006. Visualization for Exploration of Uncertainty Related to Fuzzy Classification. In: *IEEE International Sensing Symposium Geoscience and Remote*. IEEE: New York.

Lunetta RS & Lyon JG (eds) 2004. *Remote sensing and GIS accuracy assessment*. Boca Raton: CRC Press.

MacEachren AM & Taylor DRF (eds) 1994. *Visualization in modern cartography*. London: Pergamon Press.

MacEachren AM 1992. Visualizing uncertain information. *Cartographic Perspectives* 13: 10-19.

MacEachren AM 1998. Visualization - Cartography for the 21st century. In: *Polish Spatial Information Association conference.* Pennsylvania State University: Pennsylvania

MacEachren AM, Robinson A, Hopper S, Gardner S, Murray R, Gahegan M & Hetzler E 2005. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science* 32, 3: 139-160.

MapCruzin s.a. Free GIS mapping, ArcGIS shapefiles, tools, news, geography maps and resources [online]. Available from: http://www.mapcruzin.com/ [Accessed 21 May 2015].

Mas JF, Filho BS, Pontius RG, Gutiérrez MF & Rodrigues H 2013. A suite of tools for ROC analysis of spatial models. *ISPRS International Journal of Geo-Information* 2, 3: 869-887.

Mashimbye ZE 2013. Remote sensing-based identification and mapping of salinised irrigated land between Upington and Keimoes along the lower orange river, South Africa. Master's thesis. Stellenbosch: Stellenbosch University, Department of Geography and Environmental Studies.

Maurya S, Ohri A & Mishra S 2015. Open source GIS: a review. In: *Proceedings of national conference on open source GIS: opportunities and challenges*. Varanasi: Indian Institute of Technology 150-155.

McDonough K 2008. ArcNews Winter 2007/2008 Issue -- Managing GIS: Growing Up GIS [online]. ESRI. Available from: http://www.esri.com/news/arcnews/winter0708articles/growing-up-gis.html [Accessed 13 May 2015].

McKeen S, Wilczak J, Grell G, Djalalova I, Peckham S, Hsie EY, Gong W, Bouchet V, Menard S, Moffet R, McHenry J, McQueen J, Tang Y, Carmichael GR, Pagowski M, Chan A, Dye T, Frost G, Lee P & Mathur R 2005. Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *Journal of Geophysical Research: Atmospheres* 110, D21.

McKenzie G, Hegarty M, Barrett T & Goodchild M 2015. Assessing the effectiveness of different visualizations for judgments of positional uncertainty. *International Journal of Geographical Information Science* 30, 2: 221-239.

McNyset K, Volk C & Jordan C 2015. Developing an Effective Model for Predicting Spatially and Temporally Continuous Stream Temperatures from Remotely Sensed Land Surface Temperatures. *Water* 7, 12: 6827-6846.

Mentaschi L, Besio G, Cassola F & Mazzino A 2013. Problems in RMSE-based wave model validations. *Ocean Modelling* 72: 53-58.

Merriam-Webster s.a. Dictionary and Thesaurus | Merriam-Webster [online]. Available from: http://www.merriam-webster.com/interstitial-ad?next=%2Fdictionary%2FAccuracy [Accessed 24 October 2015].

Miller S & Childers D 2004. *Probability and random processes*. Amsterdam: Elsevier Academic Press.

Mitchell A 1999. *The ESRI guide to GIS analysis, volume 1: Geographic patterns and relationships*. Redlands: ESRI.

Monmonier M 2006. Cartography: Uncertainty, interventions, and dynamic display. *Progress in Human Geography* 30, 3: 373-381.

Mowrer HT 2000. Uncertainty in natural resource decision support systems: sources, interpretation, and importance. *Computers and Electronics in Agriculture* 27: 139-154.

NEDARC s.a. NEDARC - Overview [online]. Available from: http://www.nedarc.org/tutorials/collectingData/index.html [Accessed 12 April 2015].

Neumann A, Freimark H & Wehrle A 2010. *Geodata Structures and Data Models*. Zürich: GeoVITe.

Nondestructive testing resource centre s.a. Uncertainty terminology [online]. Available from: https://www.nde-ed.org/GeneralResources/ErrorAnalysis/UncertaintyTerms.htm [Accessed 9 January 2015].

North MA 2009. A method for implementing a statistically significant number of data classes in the jenks algorithm. In: *Sixth International Conference on Fuzzy Systems and Knowledge Discovery*. Stevenage: IEEE Xplore.

Nulty D 2008. The adequacy of response rates to online and paper surveys: what can be done?. *Assessment & Evaluation in Higher Education* 33, 3: 301-314.

Olson JM & Brewer CA 1997. An evaluation of color selections to accommodate map users with color-vision impairments. *Annals of the Association of American Geographers* 87, 1: 103-134.

Pebesma EJ, De Jong K & Briggs D 2007. Interactive visualization of uncertain spatial and spatio-temporal data under different scenarios: An air quality example. *International Journal of Geographical Information Science* 21, 5: 515-527.

Perer A & Shneiderman B 2009. Integrating statistics and visualization: Case studies of gaining clarity during exploratory data analysis. *IEEE Computer Graphics and Applications* 3, 29: 39-51.

Petrasova A, Harmon B, Petras V & Mitasova H 2014. GIS-based environmental modeling with tangible interaction and dynamic visualization. In: *7th International Congress on Environmental Modelling and Software.* The International Environmental Modelling & Software Society: Utah.

PLATO 2015. *Guidelines for GISc April 2015*. Rosherville: PLATO.

Polaris Marketing Research 2012. *Four Survey Methodologies: A Comparison of Pros & Cons*. Survey Methods White Paper Series. Atlanta: Polaris Marketing Research 1-7.

Pontius RG & Millones M 2011. Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing* 32, 15: 4407-4429.

QGIS s.a.b. QGIS API documentation: QGIS [online]. Available from: http://qgis.org/ [Accessed 2 June 2016].

QGIS s.a.b. QGIS API documentation: QGIS [online]. Available from: http://qgis.org/api/ [Accessed 11 June 2016].

Repas M 2010. *Using free, open-source software in local governments: streamlined internal computing for better performance and record keeping*. Washington: ICMA.

Rethman, C. (2014). Registration of Professional GIS Practitioners in South Africa [online]. Wahenga. Available from: http://wahenga.co.za/index.php/item/34-registration-of-professional-gis-practitioners/34-registration-of-professional-gis-practitioners [Accessed 6 May 2016].

Rogelberg S & Stanton J 2007. Introduction: Understanding and dealing with organizational survey nonresponse. *Organizational Research Methods* 10, 2: 195-209.

Rogerson P 2001. *Statistical methods for geography*. London: Sage.

Sacha D, Senaratne H, Kwon BC, Ellis G & Keim DA 2016. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 22, 1: 240-249.

Savage NH, Agnew P, Davis LS, Ordóñez C, Thorpe R, Johnson CE, O'Connor FM & Dalvi M 2013. Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model: Description and initial evaluation. *Geoscientific Model Development* 6, 2: 353-372.

Senaratne H & Gerharz L 2011. An assessment and categorisation of quantitative uncertainty visualisation methods. In: *The 14th AGILE International Conference on Geographic Information Science*. Utrecht University: Utrecht.

Senaratne H, Gerharz L, Pebesma E & Schwering A 2012. Usability of spatio-temporal uncertainty visualisation methods. *Bridging the Geographic Information Sciences* 3-23.

Seo S 2006. A review and comparison of methods for detecting outliers in univariate data sets. Master's thesis. University of Pittsburgh.

Fowler J 2011. *Cartographic communication of point level untertainty.* Master's thesis. Columbia: University of South Carolina.

Shekhar S & Xiong H 2007. *Encyclopaedia of GIS*. Berlin: Springer Science and Business.

Shi W 2010. *Principles of modeling uncertainties in spatial data and spatial analyses*. CRC Press/Taylor & Francis: Boca Raton.

Shi W, Fisher P & Goodchild MF 2002. *Spatial data quality*. Taylor & Francis: London.

Shin M, Campbell J & Burkhart N 2016. *Essentials of Geographic Information Systems*. 2nd ed. Boston: Flat World Education.

Skeels M, Lee B, Smith G & Robertson GG 2009. Revealing uncertainty for information visualization. *Information Visualization* 9, 1: 70-81.

Slocum TA, Cliburn DC, Feddema JJ & Miller JR 2003. Evaluating the Usability of a Tool for Visualizing the Uncertainty of the Future Global Water Balance. *Cartography and Geographic Information Science* 30, 4: 299-317.

Slocum TA, MacMaster RB, Kessler FC & Howard HH 2013. *Thematic cartography and geovisualization.* 3rd ed. Pearson: London.

Smits PC, Dellepiane SG & Schowengerdt RA 2010. Quality assessment of image classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach. *International Journal of Remote Sensing* 20, 8: 1461-1486.

Sui DZ 2003. Tobler's first law of geography: A big idea for a small world? *Annals of the Association of American Geographers* 94, 2: 269-277.

Technische Universität München s.a. *Professorenprofile: Westermann Rüdiger* [online]. Available from: http://www.professoren.tum.de/en/westermann-ruediger/ [Accessed 10 April 2015].

Tegtmeier W, Hack R, Zlatanova S & Van Oosterom P 2007. Identifying the problem of uncertainty determination and communication in infrastructural development, In: Stein & Demirel (Eds), *Proceedings of the 5th International Symposium on Spatial Data Quality, June 2007*, Enschede, CDROM, 8 p.

Thomson J, Hetzler B, MacEachren AM, Gahegan M & Pavel M 2005. A typology for visualizing uncertainty. In: *Symposium on Electronic Imaging*. PennState University: Pennsylvania.

TUFTS University 2012. *Delineating watersheds from a digital elevation model*. Massachusetts: TUFTS University.

UncertWeb s.a. Uncertainty enabled web model [online]. Available from: http://www.uncertweb.org [Accessed 22 April 2015].

United States Geological Survey 2014. USGS-ASPRS efforts to quantify the quality of LiDAR data [online]. Available from: https://calval.cr.usgs.gov/wordpress/wp-content/uploads/LidarPoster-FINAL-VISID2.pdf [Accessed 1 May 2016].

University of Surrey s.a. Formulae for the standard deviation [online]. Available from: http://libweb.surrey.ac.uk/library/skills/Number%20Skills%20Leicester/page_19.htm [Accessed 8 May 2016].

Van den Berg EC, Plarre C, Van den Berg HM & Thompson MW 2008. The South African national land-cover 2000. Agricultural Research Council Institute for Soil, Climate and Water. Pretoria. (Report No. GW/A/2008/86).

Van Niekerk A 2012. Developing a very high resolution DEM of South Africa. Position IT Nov-Dec 55-60. [online]. Available from: http://www.smc-synergy.co.za/downloads/PositionIT_Nov2012%20Dem.pdf [Accessed 21 April 2015].

Van Niekerk A 2014. *Stellenbosch University digital elevation model 2013 edition*. Stellenbosch: Stellenbosch University.

Van Niekerk A 2016. *Stellenbosch University digital elevation model 2016 edition*. Stellenbosch: Stellenbosch University.

Van Oort P 2005. Spatial data quality: From description to application. Delft: Netherlands Geodetic Commission.Walford N 2011. Practical statistics for geographers and earth scientists. New York City: Wiley.

Weng Q 2012. Quantifying uncertainty of digital elevation models derived from topographic maps. In: *ISPRS Congress Melbourne 2012*. Paris: International Society for Photogrammetry and Remote Sensing.

Westra E 2014. *Building mapping applications with QGIS*. Birmingham: Packts Publishing.

Willmott CJ & Matsuura K 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 30: 79-82.

Willmott CJ, Matsuura K & Robeson SM 2009. Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment* 43, 3: 749-752.

Wittenbrink CM, Saxon E, Furman JJ, Pang A & Lodha S 1996. Glyphs for visualizing uncertainty in environment vector fields. *IEEE Transactions on Visualization & Computer Graphics* 2, 3: 266-279.

Wong DW & Sun M 2013. Handling data quality information of survey data in GIS: A case of using the American community survey data. *Spatial demography* 1, 1: 3-16.

Yaffee R & McGee M 2000. *Introduction to time series analysis and forecasting with applications of SAS and SPSS*. San Diego: Academic Press.

Yuan M 1996. Modeling semantical, temporal, and spatial information in geographic information systems, in: Craglia M, & Couclelis H (eds) *Geographic information research*—Bridging the Atlantic: London, Taylor & Francis 334–347.

Zandbergen A 2011. Error propagation modeling for terrain analysis using dynamic simulation tools in ArcGIS modelbuilder. In: *Geomorphometry 2011*. California: Geomorphometry 57-60.

Zhang J & Goodchild M 2002. *Uncertainty in geographical information.* New York: Taylor & Francis.

Zhao Z, Benoy G, Chow TL, Rees HW, Daigle JL & Meng FR 2009. Impacts of accuracy and resolution of conventional and LiDAR based DEMs on parameters used in hydrologic modeling. *Water Resources Management* 24, 7: 1363-1380.

APPENDICES

APPENDIX A ETHICAL CLEARANCE FOR CHAPTER 3

Ethical clearance for Chapter 3 survey.

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

Approval Notice
New Application

05-Aug-2015
Christ, Sven SI

Proposal #: SU-HSD-000399
Title: Visualization of uncertainty in environmental data

Dear Mr Sven Christ,

Your **New Application** received on **02-Jul-2015**, was reviewed
Please note the following information about your approved research proposal:

Proposal Approval Period: **30-Jul-2015 -29-Jul-2016**

General comments:
The researcher is reminded to obtain institutional permission from organisations as indicated in his application for ethics clearance.

Please take note of the general Investigator Responsibilities attached to this letter. You may commence with your research after complying fully with these guidelines.

Please remember to use your **proposal number** (**SU-HSD-000399**) on any documents or correspondence with the REC concerning your research proposal.

Please note that the REC has the prerogative and authority to ask further questions, seek additional information, require further modifications, or monitor the conduct of your research and the consent process.

Also note that a progress report should be submitted to the Committee before the approval period has expired if a continuation is required. The Committee will then consider the continuation of the project for a further year (if necessary).

This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki and the Guidelines for Ethical Research: Principles Structures and Processes 2004 (Department of Health). Annually a number of projects may be selected randomly for an external audit.

National Health Research Ethics Committee (NHREC) registration number REC-050411-032.

We wish you the best as you conduct your research.

If you have any questions or need further help, please contact the REC office at 218089183.

Included Documents:
DESC Report - Hunter, Lauren
REC: Humanities New Application

Sincerely,

Clarissa Graham
REC Coordinator
Research Ethics Committee: Human Research (Humanities)

APPENDIX B CHAPTER 3 SURVEY

Survey for Chapter 3.

Form Title

By clicking accept I agree to take part in a research study entitled Uncertainty visualization in environmental data and conducted by Sven Christ. I declare that: • I have read the Above information and it is written in a language with which I am fluent and comfortable. • I have been given the opportunity to enquire about the project and questionnaire. • I understand that taking part in this study is voluntary and I have not been pressurised to take part. • I may choose to leave the study at any time and will not be penalised or prejudiced in any way. • All issues related to privacy and the confidentiality and use of the information I provide have been explained to my satisfaction.*

- I accept
- I decline

Add item

After page 1

Continue to next page

Page 2 of 9

1. Age

- ≤19
- 20-29
- 30-39
- 40-49
- 50-59
- 60+

2. Sex

- Male
- Female

Add item

After page 2

Continue to next page

Page 3 of 9

3. How long have you been working with geospatial data? (In Years)

4. How often do you use geospatial data?

- Daily

- Weekly
- Monthly
- Other:

5. What do you use geospatial information for?

- Developing datasets
- Analysis
- Decision making
- Other:

6. What type of decisions do you make based on geospatial information?

- Policy
- Research
- Land use/Development
- Other:

7. Are you aware of the inherent uncertainty of geospatial data?

- Yes
- No

8. Do you ask for the accuracy rating of data that you use?

- Yes
- No

9. How do you feel about a map with 80% accuracy?

10. How do you deal with uncertainty in geospatial data?

Add item

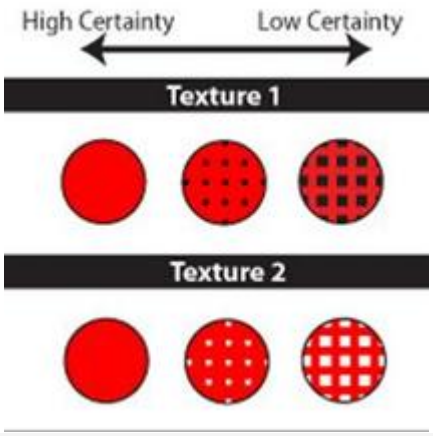
After page 3

Continue to next page

Page 4 of 9

11. How easy is the following visualization of uncertainty to interpret?

	1	2	3	4	5	
Very easy						Very difficult



Add item

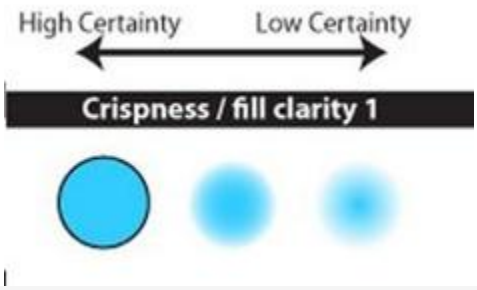
After page 4

Continue to next page

Page 5 of 9

12. How easy is the following visualization of uncertainty to interpret?

	1	2	3	4	5	
Very easy	○	○	○	○	○	Very difficult



Add item

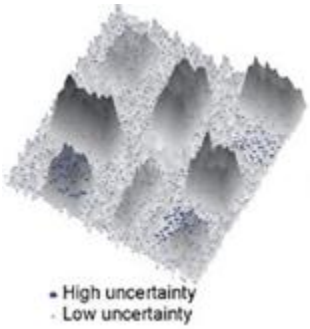
After page 5

Continue to next page

Page 6 of 9

13. How easy is the following visualization of uncertainty to interpret?

	1	2	3	4	5	
Very easy	○	○	○	○	○	Very difficult



Add item

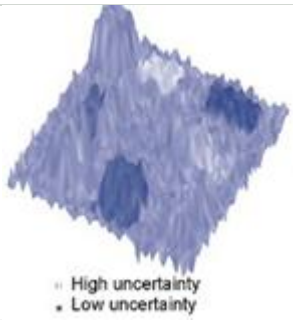
After page 6

Continue to next page

Page 7 of 9

14. How easy is the following visualization of uncertainty to interpret?

	1	2	3	4	5	
Very easy	○	○	○	○	○	Very difficult



Add item

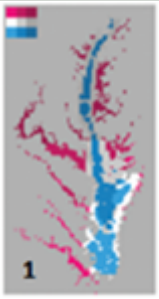
After page 7

Continue to next page

Page 8 of 9

15. How easy is the following visualization of uncertainty to interpret?

	1	2	3	4	5	
Very easy	○	○	○	○	○	Very difficult



Blue= High certainty

Red= Low certainty

Add item

After page 8

Continue to next page

Page 9 of 9

16. Have you ever come into contact with maps with visual uncertainty representations?

- Yes
- No

17. If yes, where and what method was employed?

18. Would you prefer visual or statistical uncertainty representation?

19. Do you visualize data with colour blind people in mind?

- Yes
- No

20. Do you have any comments or concerns about uncertainty visualization?

Add item

Confirmation Page

Show link to submit another response
Publish and show a public link to form results
Allow responders to edit responses after submitting

Send form

APPENDIX C ETHICAL CLEARANCE FOR CHAPTER 6

Ethical clearance for Chapter 6 interviews.

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

Approval Notice
Amendment

05-Apr-2016
Christ, Sven SI

Proposal #: SU-HSD-000399
Title: Visualization of uncertainty in environmental data

Dear Mr Sven Christ,

Your **Amendment** received on **15-Feb-2016**, was reviewed by members of the **Research Ethics Committee: Human Research (Humanities)** via Expedited review procedures on **31-Mar-2016** and was approved.
Sincerely,

Clarissa Graham
REC Coordinator
Research Ethics Committee: Human Research (Humanities)

APPENDIX D INFORMED CONSENT AND DISCUSSION GUIDE CHAPTER 6

This appendix contains the ethical documents and the interview guide for the focussed interviews in Chapter 6.

STELLENBOSCH UNIVERSITY

CONSENT TO PARTICIPATE IN RESEARCH

Visualization of uncertainty in environmental data

You are asked to participate in a research study conducted by Sven Christ (B.A. Hons Geography and Environmental Studies), from the Department of Geography and Environmental Studies at Stellenbosch University. The results will form part of a Master's Thesis in the Department of Geography and Environmental Studies. You were selected as a possible participant in this study because you are seen as a person who is involved with the creation as well as use of geospatial information.

1. PURPOSE OF THE STUDY

The study aims to assess what the level of knowledge is amongst those in the geospatial industry as well as introducing a tool to visualize geospatial uncertainty. This interview will form part of the evaluation of the created tool if it is useful and what shortcomings it may have.

2. PROCEDURES

If you volunteer to participate in this study, we would ask you to do the following things:

Agree to an interview at a time and location convenient to the interviewee.

In the interview answer a few basic questions about the understanding of uncertainty in geospatial data.

Watch a demonstration of uncertainty visualization with the software tool that has been created in this study. The data will be preselected by the researcher and as such holds no influence on any work you have done or may in the future do.

Answer some questions about how you feel about the data before and after the demonstration as well as about the usability of the tool.

3. POTENTIAL RISKS AND DISCOMFORTS

There are no foreseeable risks or discomforts that an interviewee may experience. If any situation should occur the interview will be cancelled and if required all data from it destroyed.

4. POTENTIAL BENEFITS TO SUBJECTS AND/OR TO SOCIETY

The benefit to the interviewee may only extend to being informed to inspecting any data they use through a different method. Namely the freely available software tool that has been developed in this study.

Potential benefits to science and education include an insight into the world of uncertainty and data quality. There are already studies that look into how uncertainty and data quality is being perceived in geospatial data. They have however all been limited studies. This research although limited as well will give a view on the situation in South Africa. It will also add to the body of knowledge as to how geospatial data is often viewed.

5. PAYMENT FOR PARTICIPATION

No payment will be given for participation in this study.

6. CONFIDENTIALITY

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law. Confidentiality will be maintained by means of being kept on a password protected computer as well as names removed from responses. No names will be used in any written document be it published only to the university library or any other publication.

If the interviewee agrees to allow an audio recording during the interview he/she shall have the right to review and edit the tapes. They will however be destroyed once the research has been submitted for review. The audio recording is however not compulsory and if an interviewee declines to have the session recorded it will not result in the interview being cancelled.

The information will only be used for this study and all raw information will only be held by the researcher.

7. PARTICIPATION AND WITHDRAWAL

You can choose whether to be in this study or not. If you volunteer to be in this study, you may withdraw at any time without consequences of any kind. You may also refuse to answer any questions you don't want to answer and still remain in the study. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

8. IDENTIFICATION OF INVESTIGATORS

If you have any questions or concerns about the research, please feel free to contact

Sven Christ (principal researcher):

Email- 16823745@sun.ac.za or 0843006232.

9. RIGHTS OF RESEARCH SUBJECTS

You may withdraw your consent at any time and discontinue participation without penalty. You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you have questions regarding your rights as a research subject, contact Ms Maléne Fouché [mfouche@sun.ac.za; 021 808 4622] at the Division for Research Development.

SIGNATURE OF RESEARCH SUBJECT OR LEGAL REPRESENTATIVE

The information above was described to [*me/the subject/the participant]* by [*name of relevant person]* in [*Afrikaans/English/Xhosa/other]* and [*I am/the subject is/the participant is]* in command of this language or it was satisfactorily translated to [*me/him/her*]. [*I/the participant/the subject]* was given the opportunity to ask questions and these questions were answered to *[my/his/her]* satisfaction.

[I hereby consent voluntarily to participate in this study/I hereby consent that the subject/participant may participate in this study.] I have been given a copy of this form.

_____

Name of Subject/Participant

_____

Name of Legal Representative (if applicable)

_____ _____

Signature of Subject/Participant or Legal Representative Date

SIGNATURE OF INVESTIGATOR

I declare that I explained the information given in this document to _____ [*name of the subject/participant]* and/or [his/her] representative _____ [*name of the representative]. [H*e/she] was encouraged and given ample time to ask me any questions. This conversation was conducted in [*Afrikaans/*English/*Xhosa/*Other*] and [*no translator was used/this conversation was translated into* _____ by _____].

_____ _____

Signature of Investigator Date

Interview Guide

Consent Process

Consent forms for focus group participants are emailed in advance to all participants so they may familiarize themselves with it. Signing will occur on arrival of the interviewer in a printed form supplied by the interviewer. The lead researcher (Sven Christ) will do all the interviews.

1. Welcome
   A short introduction of the researcher and the purpose of the study. Any clarity needed on the study will also be given in interviewees requests.

2. Explanation of the process
   Explain why the participant was chosen and how the data will be used.

- In this project, we are doing both questionnaires and focus group discussions. The reason for using both of these tools is that we can get more in-depth information from a smaller group of people in focus groups. This allows us to understand the context behind the answers given in the survey and helps us explore topics in more detail than we did in the survey.

Logistics

- Focus group will last about one hour
- Feel free to move around

3. Ground Rules

   Ask the participants if there are any ground rules that they would like to include to the discussion.

4. Turn on Tape Recorder

   a. It the participants agree to this else the discussion will be recorded in a reviewable written format.

5. Ask the participant if there are any questions before we get started, and address those questions.

6. Introductions
   - Go around table: job here, where you were born

Discussion begins, make sure to give people time to think before answering the questions and don't move too quickly. Use the probes to make sure that all issues are addressed, but move on when you feel you are starting to hear repetitive information.

Semi-structured question themes:

1. What is understood by uncertainty in geospatial data?

2. How is uncertainty managed, would visualization aid in this? (internally)

3. Discussion about dataset quality in statistical terms

   Do demonstration of the developed uncertainty visualization tool

4. How do you feel about the data now? (does the visualization degrade the perceived quality of the data)

5. Would visualization of uncertainty aid in this management?

6. How is uncertainty communicated to end users and would visualization aid in this?

7. Do you feel the visualization of uncertainty would degrade the perceived value of a dataset?

8. Are there any suggestion or remarks about the usability of the developed tool?

That concludes our interview. Thank you so much for coming and sharing your thoughts and opinions with us

Materials and supplies for focus groups

- Consent forms (one copy for participants, one copy for the researcher)
- Discussion Guide for Researcher
- 1 recording device
- Notebook and Pen for note-taking
- Refreshments